# Evaluation of BUSOGO climate Model via residuals for one decade

A. Nkurunziza,[1,*] J. Nzabanita,[1] and W. Banzi[1]

[1] *Department of Mathematics, School of Science, University of Rwanda, P.O Box 3900, Kigali, Rwanda.*

Testing how a chosen model fits the data and checking the effectiveness of the present model is important in statistics to be efficient. In this paper, the analysis of residuals was used to evaluate BUSOGO Climate Statistical Model. Some variable selection method (stepwise selection) has been used to decide how the rainfall can be predicted at BUSOGO region. The predicted values ($Y_i$) were found and a residual plot was constructed against the fitted values, and the assumptions made for the selected model were verified and outliers were identified. Residual plots were investigated for setting up a mean shift model for each observation and the distribution and independence of $R_i$ was assessed. Furthermore, $n$ statistic tests ($T_i$) about $\mu$ for each residual test were conducted. by using null hypothesis ($H_0 : \mu = 0$) against alternative hypothesis ($H_1 : \mu \neq 0$). Findings showed that the glass minimum temperature and the relative humidity are the major attributes to predict rainfall at BUSOGO region. The test of normality shows that the observed sample comes from a distribution which is normal. Outliers analysis showed that null hypothesis ($\mu = 0$) is accepted. This implies that there are no outliers in the selected model.

* nkuralex2018@yahoo.com

## 1. Introduction

In statistical analysis, one of the main objectives is to test how a chosen model fits the data and check whether these data fulfill the chosen model assumptions. However, there exist many ways of doing this and different statistical tests for evaluating the model assumptions. One of these ways is the so called analysis of residuals where the residuals are defined to be the deviation from the observed value to the predicted ones in the model. Thus, by looking carefully at the residuals one can observe whether the assumptions made are reasonable and of course if the choice of the model is appropriate. In this work we will proceed as said above and investigate the fitness of a selected model (Busogo climate model) for made assumptions by studying different residual plots and there after perform the analysis for patterns and trends.

however, at the ISAE-BUSOGO station, recording rainfall data on daily basis is particulary important. This is very laborious and makes the system less predictable for the future; resulting in lack of planning for farmers for example in cropping periods. Besides, no available predicting model that has been developed at the present station until today. Other major factors include the role of every parameter involved in the model, interpretation of coefficients, etc. In this paper the main problem is to find the "best" subset of independent variables in the model

$$Y = X\beta + \varepsilon \tag{1}$$

using one (or more) of the model selection criteria; where $Y$ is a $n \times 1$ vector of observations of the response variable, $\beta$ is a $p \times 1$ vector of parameters to estimate, $\varepsilon$ is a $n \times 1$ vector of random errors and $X$ is a fixed $n \times p$ design matrix.

Then, the development of the present model will proceed by different steps, such as proposing a regression equation, model diagnostic, results understanding, etc. Therefore we want to use numerical or graphical methods to analyze the assumptions. The statistical method commonly used is the residual analysis.

Therefore, the rainfall is a major characteristic of climate system, and is used to determine climate fluctuations due to having a significant influence on ecosystem. Latitude, altitude, flora, and landscape are the main factors of climate variations from one place to another. Other climate variations include seasonal, annual, centurial, or even for much longer time. Climate change or climate variability refers to significant variations of the mean of the climate data over at least a decade. In this paper, the broad objective is to develop a predicting model of monthly rainfall at BUSOGO region based on daily records on climatic elements and to perform a residual analysis. The variables are: $X_1$ : temperature under Stevensoon (Celsius degree), $X_2$ : glass minimum temperature (Celsius degree), $X_3$ : maximum temperature (Celsius degree), $X_4$ : minimum temperature (Celsius degree ), $X_5$ : soil temperature on 10cm (Celsius degree), $X_6$ : soil temperature on 20cm (Celsius degree), $X_7$ : cloud cover (in octas), $X_8$ : sunshine (in hours)/day, $X_9$ : evaporation under Stevensoon screen (in $mm$), $X_{10}$ : back evaporometer (in $mm$), $X_{11}$ : relative humidity (in percentage), $X_{12}$ : vapour pressure (in $hPa$) and $Y$ : precipitation (Rainfall) (in $mm$)/day.

## 2. Methodology

In order to achieve our objectives we will proceed as follows:

(i) Use some variable selection method to decide how the rainfall can be predicted.

(ii) Find out the predicted value $\hat{Y}_i$.

(iii) Calculate the residuals $R_i = Y_i - \hat{Y}_i$, where $Y_i$ is $n \times 1$ vector of observations of the response variable,

(iv) Construct a residual plot against the fitted value, verify the assumptions made for the selected model and identify outliers.

(v) By investigating the residual plots, set up a mean shift model for each observation: $R_i = \mu z_i + \varepsilon$, where $z_i = 1$ for the $i^{th}$ observation and 0 elsewhere. Investigate the distribution of $\varepsilon$ and check the independence of $R_i$.

(vi) Construct $n$ test statistics $T_i$ about $\mu$, such that for each residual test via $T_i$ the hypothesis is

$$H_0 : \mu = 0$$

against

$$H_1 : \mu \neq 0.$$

The present study will focus on two aspects. The first will be the regression diagnostic by using residual analysis through statistical tests and plotting. Another aspect is model improvement by variable selection. For the fitted regression model, the relationship between each predictor and the response variable is approximately linear. This assertion is verified by drawing a scatter plot and checking if predictors and the response variable are linked by a linear relationship. The assumptions related to residuals are highlighted as follows. The residuals are:

- uncorrelated and
- normally distributed with mean zero; i.e $E(\varepsilon) = 0$ and constant variance $\sigma^2$.

From the summary statistics, such as *SSR* (Sum of square of regression), *SSRes* (Sum of square of residuals), t-and F-test statistics or $R^2$ (coefficient of determination) it can be very hard (or impossible) to tell whether the assumptions are satisfied or violated.

### 2.1. Matrix Form of Multiple Linear Regression

Let $Y$ denote the dependent variable that is linearly related to $p-1$ independent (or explanatory) variables $X$'s through the parameters $\beta$'s and write the corresponding model as follows:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i. \tag{2}$$

The model above is called multiple linear regression model, where $i = 1, \cdots, n$. The parameters $\beta_i$'s are the regression coefficients associated with $X_{i,p-1}$'s respectively and $\varepsilon$ is the error vector. The matrix form of this model is of the form:

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} 1 & X_{1,1} & X_{1,2} & \dots & X_{1,p-1} \\ 1 & X_{2,1} & X_{2,2} & \dots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \dots & X_{n,p-1} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\varepsilon}. \qquad [?]$$

It is clear that our model has the matrix form

$$Y = X\beta + \varepsilon. \tag{3}$$

### 2.2. Estimation

Consider the model 3 and the assumptions made here above and let $B$ be the set of all possible vectors $\beta$. Note that $\varepsilon = Y - X\beta$. The sum of squared residuals [**?** ] is

$$SS_{Res} = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta). \tag{4}$$

The aim is to find a vector $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_n)$ from $B$ that minimizes 4. Let us rewrite 4 as

$$SS_{Res} = Y'Y + \beta' X'X\beta - 2\beta' X'Y. \tag{5}$$

To find the value of $\beta$ that minimizes 5, we differentiate 4 with respect to $\beta$. We obtain

$$\frac{\partial SS_{Res}}{\partial \beta} = 2X'X\beta - 2X'Y.$$

Set the first derivative to zero (first order condition) yields what are called *the normal equations*

$$X'X\hat{\beta} = X'Y.$$

If $X$ is of full rank $p$, then $X'X$ is positive definite and we obtain a unique solution

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{6}$$

where $\hat{\beta}$ is a minimizer of

$$SS_{Res} = (Y - X\beta)'(Y - X\beta),$$

the sum square of residuals ($SS_{Res}$) [**? ?** ]. Thus we have the estimator $\hat{Y}$ of $Y$ given by

$$\hat{Y} = X\hat{\beta}. \tag{7}$$

### 2.3. Prediction Matrix

One can see that $\hat{Y} = Xb = X(X'X)^{-1}X'Y = PY$, where

$$P = X(X'X)^{-1}X' \tag{8}$$

is called the prediction matrix (Hat matrix) which is symmetric and idempotent. We call this "Hat matrix", because it turns $Y$'s into $\hat{Y}$. Hat matrix relates the fitted values to the observed values. It describes the influence each observed value has on each fitted value and contains useful information for detecting outliers and identifying influential observations.

### 2.4. Residuals

From the defined prediction matrix 8, we define the residual vector as

$$\begin{aligned} R &= Y - X\hat{\beta} \\ &= Y - X(X'X)^{-1}X'Y \\ &= [I_n - X(X'X)^{-1}X']Y \\ &= MY \end{aligned}$$

where $I_n$ denotes the identity matrix and $M = I_n - X(X'X)^{-1}X'$ is a symmetric idempotent matrix. For idempotent matrices the rank is equal to the trace. Substituting for $Y$ we obtain:

$$\begin{aligned} R &= M(X\beta + \varepsilon) \\ &= MX\beta + M\varepsilon \end{aligned}$$

Since

$$\begin{aligned} MX &= (I_n - X(X'X)^{-1}X')X \\ &= X - X \\ &= 0 \end{aligned}$$

Thus

$$R = (I - P)\varepsilon \qquad [?] \tag{9}$$

### 2.5. Variable selection in multiple linear regression

Regression analysis has three major applications: *description, control* and *prediction.*

The method of sustainable regressions is infeasible when the number of predictors is large. A common alternative method in this case is focused on applying a stepwise algorithm. There exist three types of stepwise procedures available: **forward addition**, **backward elimination** and **stepwise search** [**? ? ?** ].

**Forward addition** begins by determining which one of the $X$-variables provides most information about $Y$. This variable is retained in all future models. At the second stage the procedure considers the remaining $(p-1)$ variables and determines which, in conjunction with the first variable, provides most additional information about $Y$. This procedure continues until there are no further variables that make useful extra contributions to the fit of the model.

**Backward elimination** represents forward selection by starting with the model containing all $X$-variables and removing ineffective variables one by one. A variable is considered to be ineffective if its contribution results in a value for the $F$-test that fails to exceed the critical value$F$ to remove in the model.

**Stepwise search** itself is more similar to the forward addition algorithm. Therefore, as in the forward addition, the most significant variable is added to the model at each step, if its corresponding $F-$test is significant at the level of $\alpha$ to enter. However, before the next variable is removed in, the stepwise search procedure takes an additional look-back step to check all variables contained in the current model and deletes any variable that has a p-value greater than $\alpha$ to stay.

### 2.6. Measures based on residuals

The residuals play an important role in regression **diagnostics**, since the $i^{th}$ residual $R_i$ may be regarded as an appropriate guess for the unknown random error $\varepsilon_i$. The relationship $R = (I - P)\varepsilon$ implies that $R$ is a good estimator for $\varepsilon$ if $(I - P) \approx I$ [**?** ], that is, if all $p_{ij}$ are sufficiently small and if the diagonal elements $p_{ii}$ are of the same size. Furthermore, even if the random errors $\varepsilon_i$ have the property $E(\varepsilon\varepsilon') = \sigma^2 I_n$, it follows that the identity $R = (I - P)\varepsilon$ shows also that the residuals are not independent (unless $P$ is diagonal) and do not have the same variance (unless the diagonal elements of $P$ are equal). Consequently, the residuals can be expected to be reasonable substitutes for the random errors if the following hold:

- the diagonal elements $p_{ii}$ of the matrix $P$ are almost equal, that is, the rows of $X$ are almost homogeneous, implying homogeneity of variances of the $\varepsilon_i$ and
- the off-diagonal elements $p_{ij}, (i \neq j)$ are sufficiently small, imply uncorrelated residuals.

We may use transformed residuals for diagnostic purposes. This means that instead of using $R_i$ we should use a transformed standardized residual, say $r_i = R_i/\hat{\sigma}_i$, where $\hat{\sigma}_i$ is the standard deviation of the $i^{th}$ residual. We can obtain several standardized residuals with specific diagnostic power according to different choices of $\hat{\sigma}_i$. For our choice, we will need the so called Studentized residual. This can be defined as externally or internally as follows.

For internally Studentized residual, with $\hat{\sigma}_i = s\sqrt{1 - p_{ii}}$, we have

$$r_i = \frac{R_i}{s\sqrt{1 - p_{ii}}}, \qquad (i = 1, \cdots, n). \qquad [?] \tag{10}$$

Note that $\sigma$ is unknown and it can be estimated by $s$, where $s$ is the mean square error of a regression model defined by

$$s = \sqrt{(R'R)/(n - p)}. \tag{11}$$

The externally Studentized residual follows by assuming that the $i^{th}$ observation is omitted. This fact will be indicated by writing the index $(i)$ in brackets. Using this indicator, we may define the estimator of $\sigma_i^2$ with the $i^{th}$ row $(Y_i, X_i)$ omitted as

$$S^2_{(i)} = \frac{Y'_{(i)}(I - P_{(i)})Y_{(i)}}{n - p - 1}, \qquad (i = 1, \cdots, n) \tag{12}$$

and by taking

$$\hat{\sigma}_i = s_{(i)}\sqrt{1 - p_{ii}},$$

we have that the $i^{th}$ externally Studentized residual is defined as

$$r_i^* = \frac{R_i}{s_{(i)}\sqrt{1 - p_{ii}}}, \qquad (i = 1, \cdots, n). \tag{13}$$

### 2.7. Linear model of residuals

After fitting a multiple regression model, and calculate the parameter estimates, we need to make analysis of the residuals in order to detect the outliers. Once the outliers are detected, we consider the so called "mean shift model for outliers". This is defined as:

$$Y = X\beta + R_i \tag{14}$$

with

$$R_i = z_i\mu + \varepsilon_i$$

where $z_i$ is the $i^{th}$ unit vector, so that $z_i' = (0, \cdots, 0, 1, 0, \cdots, 0)$. This model is known as the linear model of residuals and it is used to test the systematic deviation between $Y_i$ and $X\beta$ from the model (1). To do this we need to test the hypothesis

$$H_0 : \mu = 0 \qquad \Longrightarrow E(Y) = X\beta$$

against the alternative

$$H_1 : \mu \neq 0 \qquad \Longrightarrow E(Y) = X\beta + z_i\mu$$

and this using the $t-$test statistic;

$$t_i = \frac{Y_i - \widehat{Y}_i}{S_{(i)}\sqrt{1-p_{ii}}} \sim t_{n-p-1} \tag{15}$$

In this case the $i^{th}$ observation is a potential mean shift outlier and then $S^2$ is a biased estimate of the error variance $\sigma^2$; therefore, to construct the above test statistic, we would prefer to use the leave-one-out $S_{(i)}$ and replacing $\sigma$ by $S_{(i)}$. This $t-$statistic is often called the $R-$student statistic. If the $i^{th}$ observation is a mean shift outlier, so the $t-$statistic tends to be larger and the rejection of $H_0$ implies that the $i^{th}$ observation may be a possible mean shift outlier [**?** ].

## 3. Results and Discussion

### 3.1. Variable selection

The data are recorded at ISAE-BUSOGO station. We will now illustrate model choice in detail by means of the introduced criteria on the basis of a data set. Consider the following model with 12 real regressors and $N = 66$ observations referring to our data set.

$$\begin{aligned} Y_i = {} & \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \beta_4 X_{i,4} + \beta_5 X_{i,5} \\ & + \beta_6 X_{i,6} + \beta_7 X_{i,7} + \beta_8 X_{i,8} + \beta_9 X_{i,9} + \beta_{10} X_{i,10} \\ & + \beta_{11} X_{i,11} + \beta_{12} X_{i,12} + \varepsilon_i, \quad i = 1, 2, \cdots, 66. \end{aligned}$$

#### *3.1.1. Procedure of stepwise selection*

Using Matlab software,

1. The stepwise procedure starts with no variables in the model and chooses the predictor with the largest correlation in absolute value with $Y$, as shown in Table 1.

TABLE 1. $X_i$ in the correlation with Y

| $X_i$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Correlation with $Y$ | 0.1252 | 0.4749 | -0.1747 | 0.4123 | 0.1791 | 0.1695 | 0.4714 | -0.2352 | -0.4902 | -0.2291 | 0.5755 | 0.3658 |

2. Test if this predictor is significant: $\beta_1 = 0$ against $\beta_1 \neq 0$.
3. Take the first chosen predictor and add one predictor at the time: (choose the one with smallest $SS_{Res}$) and test if this predictor is also significant: ($\beta_2 = 0$ against $\beta_2 \neq 0$).
4. Repeat 3).

For our case, stepwise procedure starts with no variable in the model and first chooses the variable $X_{11}$: **relative humidity** ( Figure 1), since $X_{11}$ shows the largest correlation with $Y$; i.e 0.5755 as shown in Table 1.

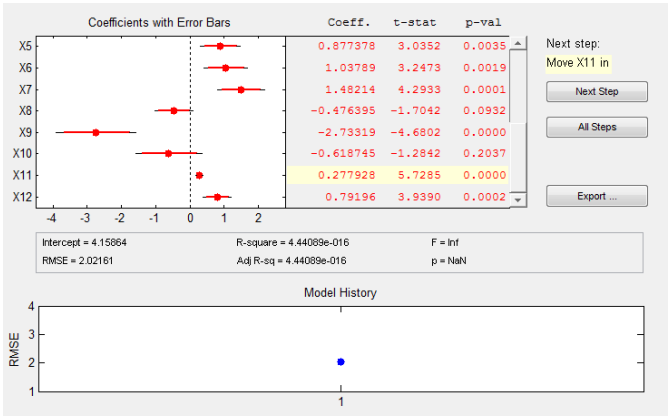



FIG. 1. Step 1: Variable Relative humidity entered

The upper left hand displays estimates of the coefficients for all potential terms, with horizontal bars indicating 90% (colored) and 95% (grey) confidence intervals. The red color shows that, initially, the terms are not in the model.

### 3.1.2. Step 2: Variable glass minimum temperature entered

The chosen next step adds the most significant term ($X_2$: Glass minimum temperature) in the model (Figure 2). At this level, the stepwise procedure computes the t-statistic for each variable currently in the model and for its estimated coefficient; squares it and describes this as its "F-to-remove" statistic. Thereafter, it computes the t-statistic for each variable not in the model that its coefficient would have if it was the next variable added; squares it and reports this as its "F-to-enter" statistic.

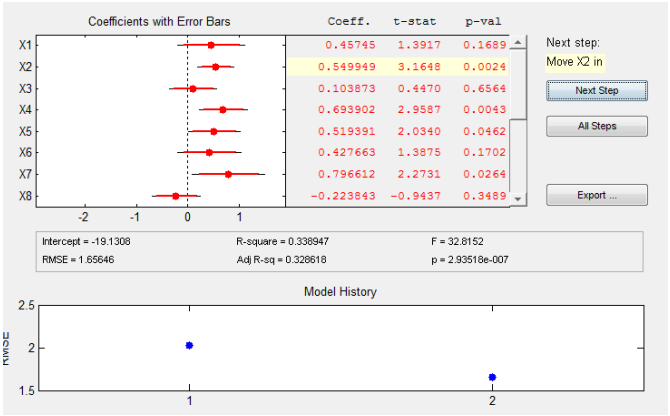



FIG. 2. Step 2: Variable glass minimum temperature entered

At the level 3, the stepwise procedure shows that there is no new predictor that can be entered in the model. It means that the predictors: $X_2$; and $X_{11}$ are added in the model whereas $X_1$, $X_3$, $X_4$, $X_5$, $X_6$, $X_7$, $X_8$, $X_9$, $X_{10}$, $X_{12}$ do not contribute significantly to the model. Therefore, those ones are eliminated from the model. As a result, our model will be:

$$\widehat{Y_i} = \beta_0 + \beta_2 X_{i,2} + \beta_{11} X_{i,11} \tag{16}$$

One can say that the rainfall at BUSOGO region depends on the glass minimum temperature and on the relative humidity. Others attributes do not contribute significantly to the model.

### 3.1.3. Model summary

The model summary in Table 2 presents two models: model 1 refers to the first stage and contains $X_{11}$ as predictor. Model 2 refers to the final model and contains $X_{11}$ and $X_2$ as predictors.

TABLE 2. Model Summary.

| Model | R | R square | Adjusted R square | Std. Error of the estimate | Change Statistics | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | R square change | F change | df1 | df2 | Sig. F change |
| 1 | $.575^a$ | .331 | .321 | 1.53829 | .331 | 31.690 | 1 | 64 | .000 |
| 2 | $.634^b$ | .401 | .382 | 1.46679 | .070 | 7.391 | 1 | 63 | .008 |

[1.] Predictors in the model: (Constant), $X_{11}$
[2.] Predictors in the model: (Constant), $X_{11}$, $X_2$

### 3.1.4. ANOVA

The ANOVA shown in Table 3 gives a formal $F$ test for the parameters effect. Therefore, we have to test the null hypothesis $H_0$: all regression coefficients are zero, against alternative hypothesis $H_1$: all regression coefficients are not zero. This statistic follows an $F$ distribution with $p$ and $N - p - 1$ degrees of freedom. For our case in the model 1, the critical $F$ value for $\alpha = 0.05$, $p = 1$, and $N - p - 1 = 64$ degree of freedom is 3.99. Since the $F$ statistic 31.697, is greater than the critical value, $H_0$ will be rejected. In the model 2, the critical $F$ value for $\alpha = 0.05$, $p = 2$, and $N - p - 1 = 63$ degree of freedom is 3.142 . Since the $F$ statistic 21.114, is greater than the critical value, we conclude that $H_0$ is rejected and we say that all regression coefficients in the model 1 and in the model 2 are significant and the model 2 improves significantly our ability to predict the rainfall.

TABLE 3. Anova.

| Model | | Sum of square | df | Mean square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 75.052 | 1 | 75.052 | 31.697 | .000[a] |
| | Residual | 151.537 | 64 | 2.368 | | |
| | Total | 226.588 | 65 | | | |
| 2 | Regression | 90.931 | 2 | 45.465 | 21.114 | .000[b] |
| | Residual | 135.658 | 63 | 2.153 | | |
| | Total | 226.588 | 65 | | | |

[1.] Predictors in the model: (Constant), $X_{11}$.
[2.] Predictors in the model: (Constant), $X_{11}$, $X_2$

### *3.1.5. Coefficients*

For all those steps, if we compare t-statistic to enter in the model with

$$t_{(\alpha,N-p-1)} = 1.6694$$

we find that t-statistic is always greater than $t_{(\alpha,N-p-1)}$, as shown in Table 4. In this case, we reject $H_0$: $\beta_i = 0$ for $i = 2, 11$ and the chosen model will be

$$\widehat{Y}_i = -16.384 + 0.446X_{i,2} + 0.202X_{i,11}.$$

Compare now the model with all variables: i.e, we test if

$$\beta_1 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{12} = 0.$$

TABLE 4. Betas.

| | | Unstandardized Coefficients | | Standardized Coefficients | | |
|---|---|---|---|---|---|---|
| Model | | B | Std. Error | Beta | t | Sig. |
| 1 | (Constant) | −17.251 | 3.782 | | −4.462 | .000 |
| | $X_{11}$ | .254 | .045 | .576 | 5.630 | .000 |
| 2 | (Constant) | −16.384 | 3.620 | | −4.526 | .000 |
| | $X_{11}$ | .202 | .047 | .459 | 4.305 | .000 |
| | $X_2$ | .446 | .164 | .289 | 2.716 | .009 |

### *3.1.6. ANOVA of the model* [2′]

Consider:

- the model [1′] :
  $Y_i = \beta_0 + \beta_2 X_{i,2} + \beta_{11} X_{i,11} + \varepsilon.$
- the model [2′] :
  $Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \beta_4 X_{i,4} + \beta_5 X_{i,5} + \beta_6 X_{i,6} + \beta_7 X_{i,7} + \beta_8 X_{i,8} + \beta_9 X_{i,9} + \beta_{10} X_{i,10} + \beta_{11} X_{i,11} + \beta_{12} X_{i,12} + \varepsilon.$

We have to test the hypothesis: $H_0$: all regression coefficients are zero (not $\beta_2$ and $\beta_{11}$); ( $H_0$ cannot be rejected at the 5% significance level), against alternative hypothesis $H_1$: all regression coefficients are not zero; (the null hypothesis can be rejected at the 5% level). The test statistic applied here is

$$W = \frac{(SSE^{(1)} - SSE^{(2)})/p}{SSE^{(2)}(/N - (k+p) - 1)} \sim F(p, N - (k+p) - 1). \qquad [?] \tag{17}$$

With $p = 10$ and $k = 2$. We reject $H_0$ if $W > C = F_{1-\alpha}(p, N - (k+p) - 1)$
According to the Table 3 and the Table 5, the equation 17 becomes

$$W = \frac{(135.658 - 123.096)/10}{123.096/63} = 0.642918.$$

As $C = F_{1-\alpha}(p, N - (k+p) - 1)$ is 0.379983, this implies that $C < W$ and we reject $H_0$. These, we conclude that all predictors (not $\beta_2$ and $\beta_{11}$) don't make a significant contribution to predict the rainfall at BUSOGO region.

TABLE 5. Anova of the model [2′]

| Model | | Sum of squares | df | Mean square | F | Sig. |
|---|---|---|---|---|---|---|
| 2′ | Regression | 103.339 | 12 | 8.612 | 3.708 | .000[a] |
| | Residual | 123.096 | 53 | 2.323 | | |
| | Total | 226.434 | 65 | | | |

### 3.2. Residual analysis

According to the Figure 3, the residuals are evenly distributed on both sides of $X-$axis. This implies that the assumptions $E(\varepsilon_i) = 0$ and constant variance $Var(\varepsilon_i) = \sigma^2$ are appropriate and the normality assumption is verified here by sketched histogram of residuals, by the normal plot of residuals and by the test of normality. After this we will test if the outliers, represented by the red bars, are significant.

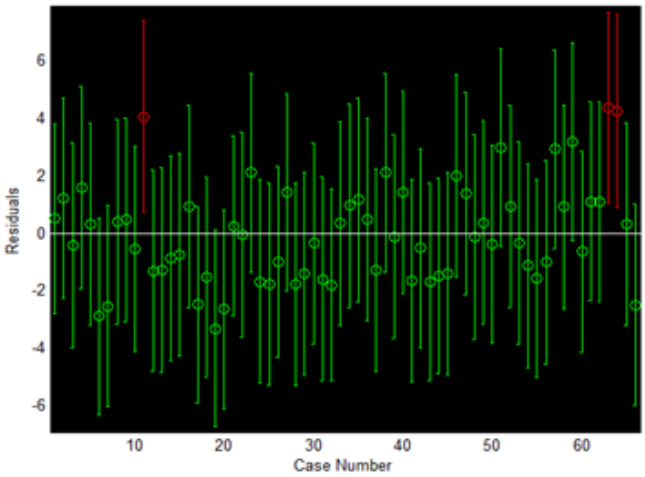
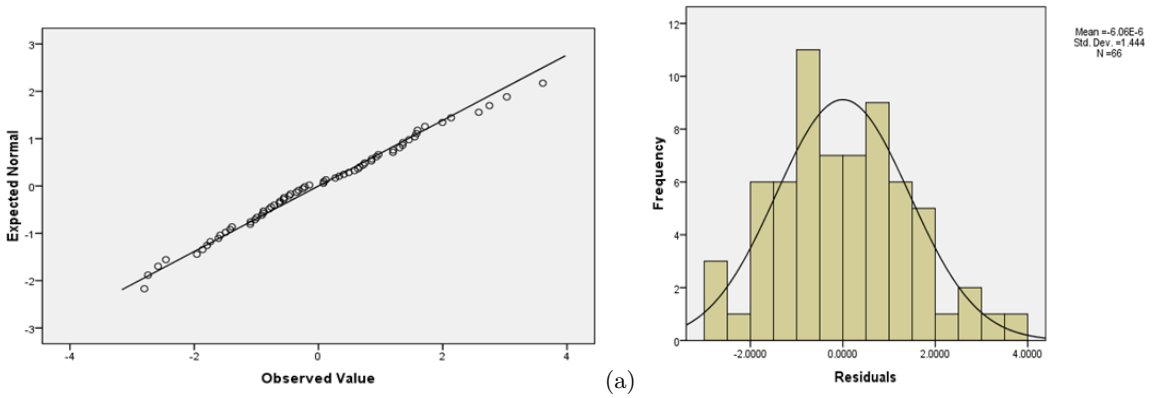



FIG. 3. Residual plot

### 3.3. Test of normality

According to the Figure 4, the distribution of residuals approximates a normal distribution. Therefore we should always check if this assumption is efficient by test.




(a) (b)

FIG. 4. (a) Normal Plot and (b) Histogram of Residuals

We know that for the data set small than 2000 elements, we use the Shapiro-Wilk test, otherwise, the Kolmogorov-Smirnov test is used [**?** ]. In our case, since we have only 66 elements, the Shapiro-Wilk test is preferred. In the last column of the Table 6, the p-value (0.780) is greater than alpha 0.05. We can reject the alternative hypothesis ($H_1$: the data are not normally distributed) and accept the null hypothesis ($H_0$: the data are normally distributed. Therefore we conclude that the data comes from a normal distribution.

TABLE 6. Test of normality.

| | Kolmogorov-Smirnov | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statics | df | Sig. | Statics | df | Sig. |
| Residuals | .061 | 66 | $.200^a$ | .988 | 66 | .780 |

### 3.4. Detection of Outliers

From the section 3.1 we have selected model containing the dependent variable $Y$ and the predictors $X_2, X_{11}$. Therefore, this model is detected to be the best compared to others. The remaining task is to identify if there are some outlier observations in the new fitted model. The calculated t-statistics $t_i$ were found to be less than $t_{\alpha,N-p-1} = t_{0.95,63} = 1.6694$. Thus, our null hypothesis $H_0 : \mu = 0$ is accepted; it implies that there are no outliers in our selected model. Next if we look at the residual plot (Figure 3) and referring to our test analysis, we conclude that the intervals shown in red are not really outliers. Hence, the residuals

$$R_i = \mu z_i + \varepsilon_i$$

are independent and $\varepsilon_i$ are normally distributed.

## 4. Conclusion

In statistical study, it is important to assess if a model under study fits the data and check whether these data fulfill the chosen model assumptions. This can be done in different ways, including analysis of residuals, and this can be done on different parameters. Climate fluctuation is one of these parameters that can be studied. One of major factors of climate fluctuation is rainfall. In this work, the major objective was to develop a predicting model of monthly rainfall at BUSOGO region based on daily records on climatic elements and to perform a residual analysis. The variables were: temperature under Stevensoon, glass minimum temperature, maximum temperature, soil temperature, cloud cover, sunshine, evaporation under Stevensoon screen, Back evaporometer, relative humidity, vapour pressure, and precipitation (Rainfall).
The stepwise selection method has been used to decide how the rainfall can be predicted at BUSOGO region. The findings from the present work show that the rainfall at BUSOGO region depends on the glass minimum temperature ($X_2$) and on the relative humidity ($X_{11}$). Therefore, the chosen model that can predict rainfall at BUSOGO region is:

$$\widehat{Y_i} = -16.384 + 0.446X_{i,2} + 0.202X_{i,11}.$$

The test of normality and outliers detection have been assessed and has shown that the data set comes from a distribution which is normal and there are no outliers in the selected model. Hence, the residuals $R_i = \mu z_i + \varepsilon_i$ is independent and $\varepsilon_i$ is normally distributed.