ADV MATH SCI JOURNAL
Advances in Mathematics: Scientific Journal **12** (2023), no.2, 329–344
ISSN: 1857-8365 (printed); 1857-8438 (electronic)
https://doi.org/10.37418/amsj.12.2.2

INFLUENCE OF UNOBSERVED DATA IN THE TIME SERIES OF THE DEPENDENT VARIABLE AND THEIR POSITION IN ANALYSIS OF MULTIPLE LINEAR REGRESSION ON PREDICTION - CASE STUDY ON: FACTORS AFFECTING CO_2 EMISSIONS

Amira I. El-Desokey

ABSTRACT. Using a variety of statistical techniques, time series forecasting is crucial for preparing for and predicting the future. It is contingent on making an accurate forecast as to the value of a variable at some unknown time in the future. This research analyzed the missing data from the time series (a model with no missing observations and three models were considered to be missing data for the dependent variable at various positions). By a standard multiple linear regression of the four models of the study, it is clear that the series is consistent, transparent, within the bounds of statistical acceptability, the analysis used the Ordinary least square and the weighted least square to find the best prediction model with missed observation.

1. INTRODUCTION

Missing observations has attracted the interest of scientists and researchers, especially missing observations in time series, which may produce less efficient estimates, some bias in the results, and inconsistency in the statistical tests used. With the development of additional statistical programs, especially in analysis,

2020 *Mathematics Subject Classification.* 62J05, 37M10.
Key words and phrases. Missing values, multiple linear regression, OLS method, WLS method, analysis of variance.
Submitted: 13.12.2022; *Accepted:* 28.12.2022; *Published:* 01.02.2023.

there has been an increase in interest in missing values, their procedures, and the method for processing them. To achieve this purpose, the researcher should pay more attention to the statistical method used, which is compatible with his data. One of the most important problems related to missing data interpolation is the estimation of future data in the presence of an imprecise history. The expected data can be evaluated using the weighted least square method. Cheng and Pourahmadi [3] proposed a technique to compute the linear regression forecasting model. This algorithm is a generalization of the innovation algorithm introduced in Brockwell and Davis [2], and it applies to any stationary time series with finite or infinite observations (1991, Prop. 5.2.2), Grenander and Rosenblatt [4], produced a closed form expression for the prediction error variance when the previous data is transformed by a finite values of missing data with arbitrary pattern. Allison [1] evaluated the effect of substituting value computation methods for treating values. According to the conclusions, linear imputation without rounding performs excellently, particularly when predicting regression coefficients in the case of simple linear regression. While Kayaalp [5] shown that the influence of missing time series observations on the parameters of the estimated model based on the least-squares method will have an effect on some of the functions the model provides. Mohsen [6] investigated the estimation and generation of data in stable time series by determining multiple forecasts using statistical software packages, as well as the possibility of utilizing some methods to find the missing values in the development of the transaction methodology. Simultaneous estimation of coefficients and missing data in time series. To estimate missing values in a time series, we have to determine out whether the missed happened on the dependent or independent side of the relationship. Only a decrease in the dependent variable is the main factor in linear regression.

The purpose of this paper is to Performing a Multiple linear regression analysis after missing some of the dependent variable observations, assessing the importance of that missed observation, as well as identifying the impact of the missed observation on the analysis. Section 2 introduces the Estimate of the model parameters for the multiple linear regressions. Section 3 presented the Weighted Least Square conditions. Section 4 introduce the practical study, we introduced

a case Study on the Co2 Mission in Egypt (1990-2019), to predict the mission in future using SPSS, conclusion, in section 5.

2. Estimate the model parameters for the multiple linear regression

We can notice that there is a clear influence during forecasting time series, especially stable ones, if the missed arises from the dependent variables during regression analysis when we evaluate the relationship between a dependent variable (Z_i) and independent multivariable (Y_i). Assuming that the independent variables are completely observed, the study will investigate missing values in the dependent variable. Some of the data were missing, and the impact of this omission will be illustrated in the multiple linear regression analysis. The analysis process will include more than one variable, one of which is dependent and the others are independent, but before we proceed, we will explore estimation methodology.

The multiple linear regression model is used to describe the relationship between the dependent variable Z and the independent variables ($y_1,\ y_2,\ \ldots,\ y_k$), which may be described by the following equation

$$Z_i = \beta_0 + \ \beta_1 y_{i1} + \beta_2 y_{i2} + \ldots + \beta_n y_{\text{in}} + \tau_i$$

$$Z_i = \beta_0 + \sum_{j=1}^{k} \beta_j y_{\text{ij}} \ + \tau_i$$

$i = 1, 2, \ldots, n$. Here $\beta_0,\ \beta_1, \beta_2, \ldots, \beta_n$the coefficients of the regression model, $\tau_i$is the random error for the observation (residuals), since we have n, observations, so we have n, equations which can be formulated as:

$$\begin{bmatrix} z_1 \\ z_2 \\ . \\ . \\ . \\ z_n \\ \ \end{bmatrix} = \begin{bmatrix} 1 & y_{11} & y_{12} \ldots y_{1k} \\ 1 & y_{21} & y_{12} \ldots y_{2k} \\ \ldots & \ldots & \ldots\ldots\ldots \\ 1 & y_{n1} & y_{n2} \ldots y_{\text{nk}} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_k \\ \ \end{bmatrix} + \begin{bmatrix} \tau_1 \\ \tau_2 \\ . \\ . \\ . \\ \tau_n \\ \ \end{bmatrix}$$

$$Z = Y\beta + V,$$

where,

- Z is the vector matrix of the dependent values,
- Y is the vector matrix of independent values,
- β is the vector matrix of the coefficient of regression,
- V is the vector matrix of random errors (residuals).

The method of ordinary least squares (OLS) is one of the most widely used approaches to estimating the parameters of a Multiple linear regression model, OLS is distinguished from other estimation techniques by the (BLUE) ((Best Linear Unbiased Estimate).

The method of ordinary least squares (OLS) is used to estimate the parameters of the multiple linear regression model, $(\beta_0,\ \beta_1, \beta_2, \ldots, \beta_n)$, which minimize the sum of the squares of the error or residual as little as possible:

$$V = Z - Y\beta$$
$$U = V^{\backslash}V = (Z - Y\beta)^{\backslash}(Z - Y\beta)$$
$$U = V^{\backslash}V = Z^{\backslash}Z - 2\beta^{\backslash}Y^{\backslash}Z + \beta^{\backslash}Y^{\backslash}Y\beta$$

In order to find the value of βthat minimizes U as little as possible, we find the derivative with respect to β_J, and set it equal to zero, giving us:

$$\frac{\partial U}{\partial \beta} = \begin{bmatrix} \frac{\partial U}{\partial \beta_0} \\ \frac{\partial U}{\partial \beta_1} \\ \ldots \\ \frac{\partial U}{\partial \beta_k} \end{bmatrix} = -2Y^{\backslash}Z + 2Y^{\backslash}Y\widehat{\beta} \tag{2.1}$$

Equation (2.1) can be reduced to:

$$Y^{\backslash}\mathrm{Y}\,\widehat{\beta} = Y^{\backslash}Z,$$
$$\widehat{\beta} = \left(Y^{\backslash}Y\right)^{-1}Y^{\backslash}Z,$$

where $\widehat{\beta}$ is the vector of the residuals.

Using equation (2.1) then we have

$$\sum_{i=1}^{n} z_i = \widehat{\beta}_0 \sum_{i=1}^{n} 1 + \widehat{\beta}_1 \sum_{i=1}^{n} y_{1i} + \widehat{\beta}_2 \sum_{i=1}^{n} y_{2i} + \ldots + \widehat{\beta}_k \sum_{i=1}^{n} y_{\mathrm{ki}},$$

$$\sum_{i=1}^{n} y_{1i} z_i = \widehat{\beta}_0 \sum_{i=1}^{n} y_{1i} + \widehat{\beta}_1 \sum_{i=1}^{n} y_{1i}^2 + \widehat{\beta}_2 \sum_{i=1}^{n} y_{1i} y_{2i} + \ldots + \widehat{\beta}_k \sum_{i=1}^{n} y_{1i} y_{\mathrm{ki}},$$

$$\sum_{i=1}^{n} y_{2i}z_i = \widehat{\beta}_0 \sum_{i=1}^{n} y_{2i} + \widehat{\beta}_1 \sum_{i=1}^{n} y_{1i}y_{2i} + \widehat{\beta}_2 \sum_{i=1}^{n} y_{2i}^2 + ... + \widehat{\beta}_k \sum_{i=1}^{n} y_{2i}y_{\text{ki}},$$

$$\sum_{i=1}^{n} y_{\text{ki}}z_i = \beta_0 \sum_{i=1}^{n} y_{\text{ki}} + \widehat{\beta}_1 \sum_{i=1}^{n} y_{1i}y_{\text{ki}} + \widehat{\beta}_2 \sum_{i=1}^{n} y_{2i}y_{\text{ki}} + ... + \widehat{\beta}_k \sum_{i=1}^{n} y_{\text{ki}}^2.$$

3. Weighted Least Square method (WLS)

In case of ordinary least square (OLS), if the dependent values don't have constant variance. If the data for the dependent variable for the regression come from a population whose distribution violates the assumption of: normality or outliers are present, and then the multiple linear regression on the original data may provide misleading results, or may not be the best test available. In such cases, fitting a different linear model or a nonlinear model, performing a weighted least squares linear regression, transforming the Y or Z data or using an alternative straight line regression method may provide a better analysis. The weighted least square linear regression (WLS) is dealing with unequal variances in Z by performing a weighted least squares fit. We use the OLS (Ordinary Least square) and WLS (Weighted least square) method to estimate the parameters $(\beta_0, \beta_1, \ldots \beta_n)$ in the multiple linear regressions.

4. Application: Study the CO_2 Mission in Egypt (1990-2019)

We investigate the effects of factors on the total mission of CO_2 including Land-Use Change and Forestry (LUCF) in Egypt from 1990 to 2019. The factors are: Building, Transportation, Land-Use Change and Forestry, in the original case (no missed observations in the dependent variable), and in three cases with missed observations in the first, middle, and end of the series. We compare the four regression models using the WLS (Weighted least squares) method and select the best fit model.

4.1. **STUDY THE NORMALITY FOR THE DATA USING SPSS.**

In Fig. 1, we examine the normality of the dependent variable and find that it does not follow the normal distribution.

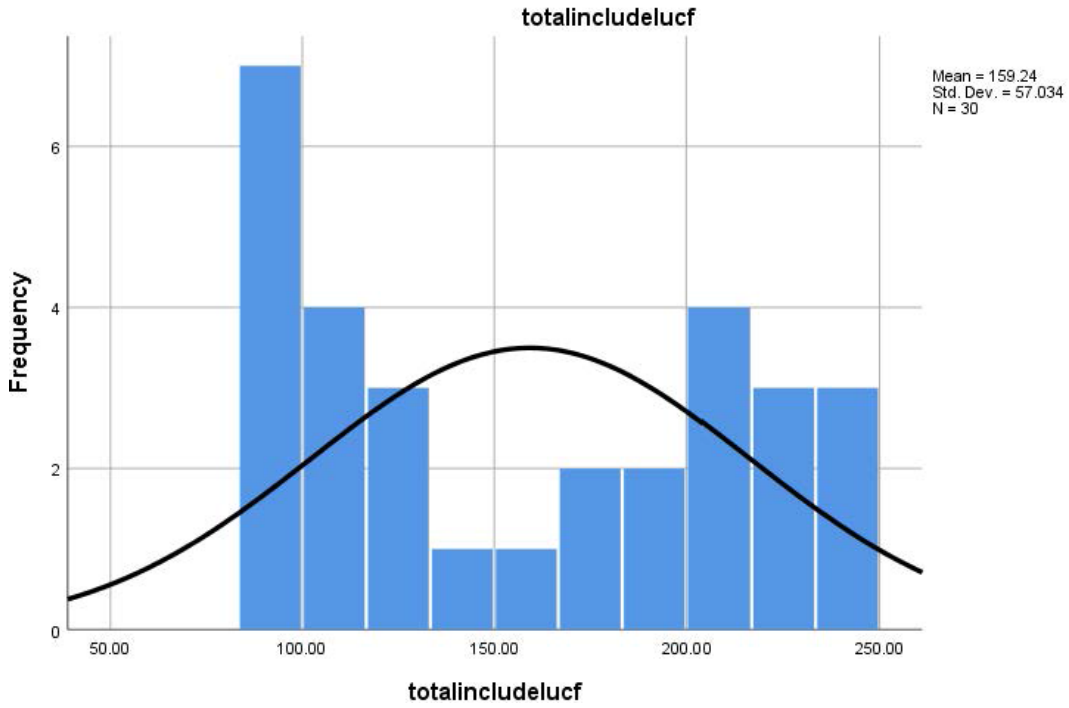

FIGURE 1.

Table 1 shows that there are unhomogeneous values of variances and that the variances are not equal.

Table 1

Descriptive Statistics						
	N	Minimum	Maximum	Mean	Std. Deviation	Variance
Total include lucf	30	87.34	249.55	159.2407	57.03442	3252.925
building	30	8.10	17.03	12.7530	2.86758	8.223
transportation	30	20.51	40.62	29.6450	5.07910	25.797
Land use change and forestry	30	-.41	0.92	-.0297	0.47246	.223
Valid N (listwise)	30					

Because of the non-normality of the data in the dependent variable, we will investigate the weighted least square multiple linear regression (WLS) instead of

the ordinary least square multiple regression (OLS) to find the best regression model.

4.2. STUDY THE WLS IN THE ORIGINAL MODEL (WITHOUT MISSED OBSERVATIONS).

We study the WLS in the original model, and Table 2, displays the Regression model summary.

Table 2

Model Summary[b,c]				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.994[a]	.988	.983	1.31058

a. Predictors: (Constant), land use change and forestry, transportation, building

b. Dependent Variable: total include lucf

c. Weighted Least Squares Regression - Weighted by weight2

We observed that the regression coefficient is 0.994[a] and the determination coefficient is 0.988.

Table 3 displays the Anova Table for the model and demonstrates that it is significant.

Table 3

ANOVA[a,b]						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2854.236	3	951.412	553.911	.000[c]
	Residual	44.658	26	1.718		
	Total	2898.895	29			

a. Dependent Variable: total include lucf

b. Weighted Least Squares Regression - Weighted by weight2

c. Predictors: (Constant), land use change and forestry, transportation, building

Table 4 discussed the model coefficients. The coefficients are β_0= -100.073, β_1=17.581, β_2=1.187, β_3= 8.449. And it shows that all the variables are significance

Table 4

Coefficients[a,b]						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-100.073	9.725		-10.291	.000
	building	17.581	.822	.879	21.380	.000
	transportation	1.187	.343	.107	3.464	.002
	Land use change and forestry	8.449	3.859	.076	2.189	.038
a. Dependent Variable: total include lucf						
b. Weighted Least Squares Regression - Weighted by weight2						

The regression model is:

$$Y_i = -100.073 + \ 17.581x_{i1} + 1.187x_{i2} + 8.449x_{i3}.$$

The time series diagram depicted in Fig. 2, demonstrates that the observed values are consistent with the fitted values.

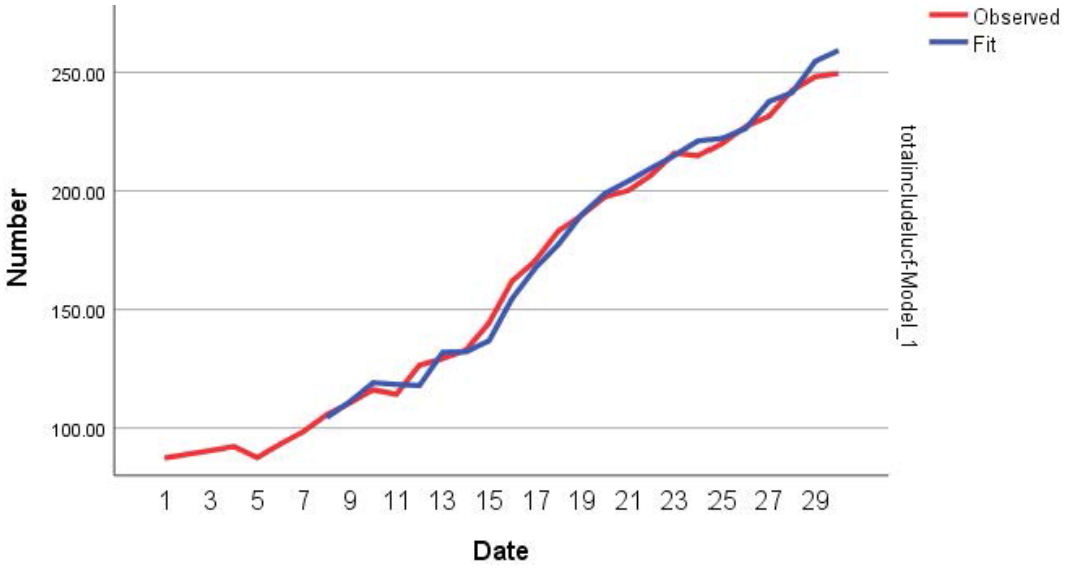

FIGURE 2.

Now we will examine the WLS method for missing observations in the dependent variable.

4.3. EXAMINES THE WLS WITH MISSED OBSERVATIONS IN THE FIRST OF THE SERIES.

Now we examine the missing observations in the first of the series to determine whether the position of the missing values has an effect on the predicted model or not. Table 5 displays a summary of the Regression model.

Table 5

Model Summary[b,c]				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
2	0.991[a]	0.983	0.980	1.29653
a. Predictors: (Constant), land use change and forestry, transportation, building				
b. Dependent Variable: total include lucf				
c. Weighted Least Squares Regression - Weighted by weight2				

We observe that the regression coefficient is 0.991[a] and the determination coefficient is 0.983.

Table 6 displays the Anova Table for the model and demonstrates that it is significant.

Table 6

ANOVA[a,b]						
Model		Sum of Squares	df	Mean Square	F	Sig.
2	Regression	2184.394	3	728.131	433.158	.000[c]
	Residual	38.663	23	1.681		
	Total	2223.057	26			
a. Dependent Variable: total include lucf						
b. Weighted Least Squares Regression - Weighted by weight2						
c. Predictors: (Constant), land use change and forestry, transportation, building						

Table 7 examined the regression model coefficients which are β_0= -99.008, β_1=17.517, β_2=1.190, β_3= 8.213. It shows that each variable has statistical significance.

Table 7

Coefficients[a,b]						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
2	(Constant)	-99.008	10.529		-9.403	.000
	building	17.517	.868	.878	20.191	.000
	transportation	1.190	.346	.117	3.436	.002
	Land use change and forestry	8.213	3.902	.079	2.105	.046
a. Dependent Variable: total include lucf						
b. Weighted Least Squares Regression - Weighted by weight2						

The regression model is:

$$Y_i = -99.008 + 17.517x_{i1} + 1.190x_{i2} + 8.213x_{i3}.$$

Fig. 3 Is a time series diagram which demonstrating that the observed values consist to the fitted values.

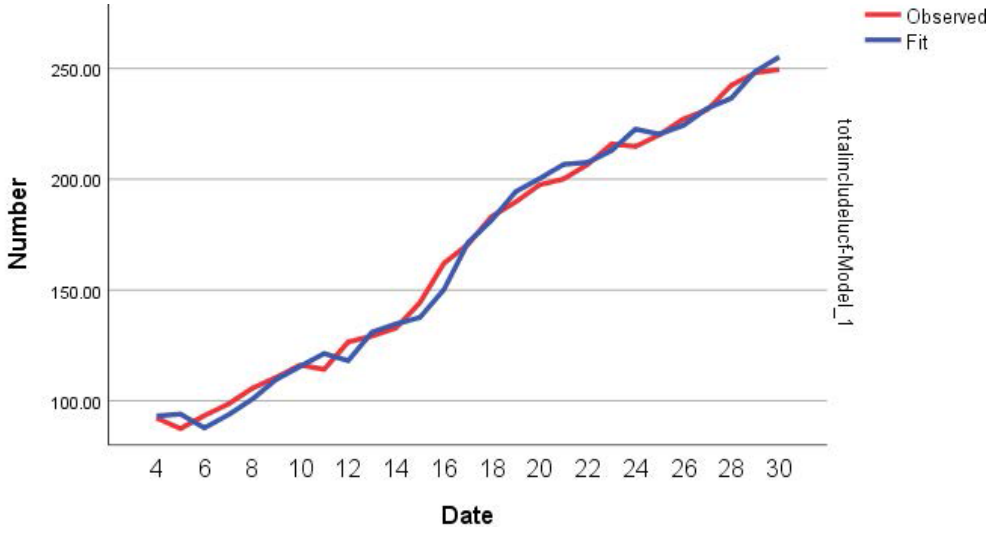

FIGURE 3.

4.4. STUDY THE WLS WITH MID-SERIES MISSED OBSERVATIONS.

Now we examine the missing observations in the mid of the series, Table 8, displays a summary of the Regression model.

Table 8

Model Summary[b,c]				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
3	0.992[a]	0.985	0.986	1.34874
a. Predictors: (Constant), land use change and forestry, transportation, building				
b. Dependent Variable: total include lucf				
c. Weighted Least Squares Regression - Weighted by weight2				

We observe that the regression coefficient is 0.992[a] and the coefficient of determination is equal 0.985. Table 9 displays the Anova Table for the model and demonstrates that it is significant.

Table 9

ANOVA[a,b]						
Model		Sum of Squares	df	Mean Square	F	Sig.
3	Regression	3399.932	3	1133.311	623.007	0.000[c]
	Residual	41.839	23	1.819		
	Total	3441.771	26			
a. Dependent Variable: total include lucf						
b. Weighted Least Squares Regression - Weighted by weight2						
c. Predictors: (Constant), land use change and forestry, transportation, building						

Table 10 examined the regression model coefficients which are β_0= -103.347, β_1=17.882, β_2=1.128, β_3= 8.233. It shows that each variable has statistical significance.

Table 10

Coefficients[a,b]						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
3	(Constant)	-103.347	8.877		-11.642	.000
	building	17.882	.837	.880	21.360	.000
	transportation	1.128	.338	.107	3.338	.003
	Land use change and forestry	8.233	3.797	.071	2.168	.041
a. Dependent Variable: total include lucf						
b. Weighted Least Squares Regression - Weighted by weight2						

The regression model is:

$$Y_i = -103.347 + 17.882x_{i1} + 1.128x_{i2} + 8.233x_{i3}.$$

Fig 4, is the time series diagram which shows that the observed values are consistent with the fit values with the declaration of the position of the missed observations.

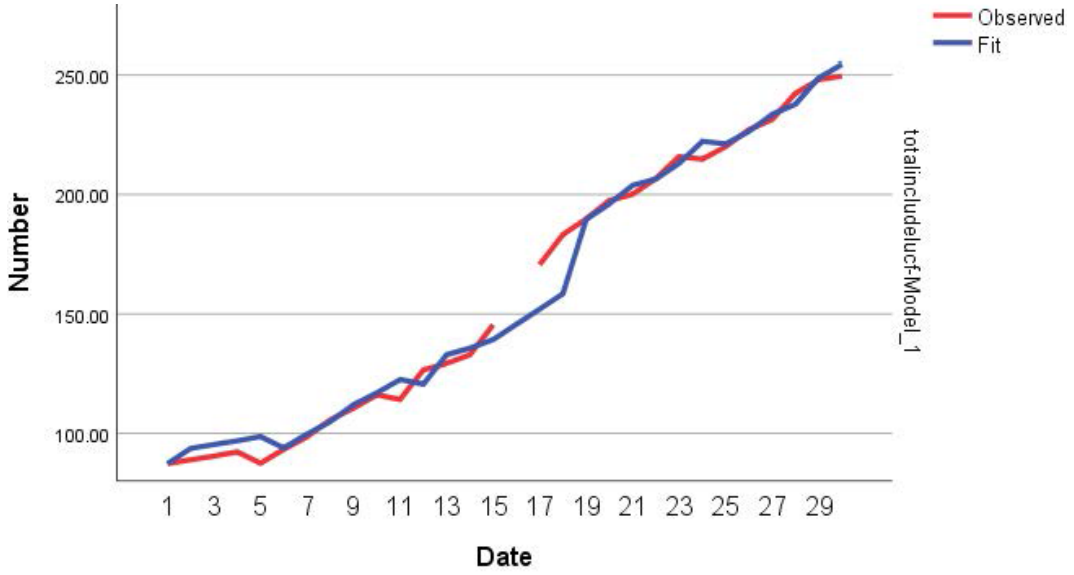

FIGURE 4.

4.5. EXAMINE THE WLS WITH MISSED OBSERVATIONS At THE END OF THE SERIES.

Now we examine the missing observations at the end of the series, Table 8, displays a summary of the Regression model.

Table 11

Model Summary[b,c]				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
4	0.990[a]	0.981	.978	1.40386
a. Predictors: (Constant), land use change and forestry, transportation, building				
b. Dependent Variable: total include lucf				
c. Weighted Least Squares Regression - Weighted by weight2				

We observed that the regression coefficient is equal 0.990[a] and the coefficient of determinant is equal 0.981.

Table 12 displays the Anova Table for the model and demonstrates that it is significant.

Table 12

ANOVA[a,b]						
Model		Sum of Squares	df	Mean Square	F	Sig.
4	Regression	2321.864	3	773.955	392.706	.000[c]
	Residual	45.329	23	1.971		
	Total	2367.193	26			
a. Dependent Variable: total include lucf						
b. Weighted Least Squares Regression - Weighted by weight2						
c. Predictors: (Constant), land use change and forestry, transportation, building						

Table 13 examined the regression model coefficients which are β_0= -91.040, β_1=17.254, β_2=1.006, β_3= 10.086. It shows that each variable has statistical significance.

Table 13

Coefficients[a,b]						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
4	(Constant)	-91.040	13.280		-6.856	.000
	building	17.254	.859	.887	20.090	.000
	transportation	1.006	.418	.076	2.404	.025
	Land use change and forestry	10.086	4.201	.100	2.401	.025
a. Dependent Variable: total include lucf						
b. Weighted Least Squares Regression - Weighted by weight2						

The regression model is:

$$Y_i = -91.040 + 17.254x_{i1} + 1.006x_{i2} + 10.086x_{i3}.$$

Fig. 5 is the time series diagram which shows that the observed values are consistent with the fit values with the declaration of the position of the missed observations.

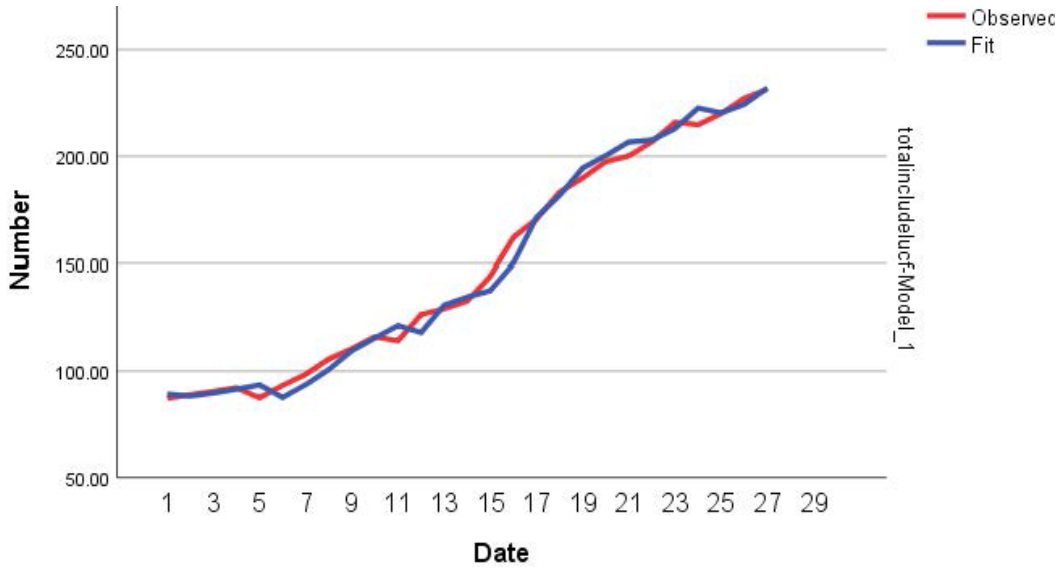

FIGURE 5.

4.6. **Comparative Analysis of Regression Models.**

As shown in Table 14, we now compare the three models of the regression with missing observations to the standard model to determine the optimal regression model.

Table 14

model	R	R^2
Standard Model	0.994	0.988
Missed observations in the first of the series	0.991	0.983
Missed observations in the mid- series	0.992	0.985
Missed observations in the end of the series	0.990[a]	0.981

The preceding table demonstrates that the model with missing observations in the mid-series, which has the highest coefficient of determination, is the best model for missing observations when compared with the standard model.

5. Conclusion

(1) If any value of the dependent variable is missed at any position, the influence of missing data time series will have an impact on the prediction of these data, as well as the degree of accuracy in the prediction.
(2) The correlation value of the four models was found to be greater than 0.90, confirming the efficacy of the data analysis method as well as the rate of missing values.
(3) it has been found that the position of the missing data significantly affects the parameter estimation, leading to a difference in the predicted values outside of the time series, regardless of whether the missing observation occurred at the beginning, the middle, or the end of the time series, particularly if the estimation is performed using the general trend approach, which is heavily impacted by the overall trend in the range of the Observations.
(4) Results from the R^2 statistic indicates that the estimated model is preferable when intermediate observations are missing from a data set. With a value of 0.985, it is the best predictive model.

References

[1] P.D. Allison: *Imputation of Categorical Variables with PROMI*, Paper presented at the Annual Meeting of the SAS users Group International, San Francisco, 2006.
[2] P.J. Brockwell, R.A. Davis: *Time Series: Theory and Methods*, Springer, 1991.
[3] Cheng and Pourahmadi: *Prediction with incomplete past and interpolation of missing values*, Statistics & Probability Letters, **33**(4) (1997), 341-346.
[4] U. Grenander, M. Rosenblatt: *An Extension of a Theorem of G. Szego and Its Application to the Study of Stochastic Processes*, Transactions of the American Mathematical Society, **76**(1) (1954), 112-126.
[5] G.T. Kayaalp: *Linear regression among independent variables in Animal breeding*, Turkish J. of Veterinary and Animal Sciences, **23**(2) (1999), 149-152.
[6] P. Mohsen: *Estimation and interpolation of missing values of a stationary time series*, Journal of time series analysis, **10**(2) (2008), 149-169.
[7] B. Pascal: *Influence of missing values on the prediction of stationary time series*, Journal of time series analysis, **26**(4) (2005), 519-525.

Departement of Computer Science
Higher Future Institute for Specialized Technological Studies
Cairo, Egypt.
Email address: aeldesokey@gmail.com, amira.eldesouky@fa-hists.edu.eg