Cairo University
Institute of Statistical Studies and Research

The 53rd Annual Conference on Statistics, Computer Sciences and Operation Research

Applied Statistics

3-5 Dec. 2018

Index
APPLIED STATISTICS

Nonstationary Time Series Analysis via Dynamic Data Systems (DDS): A New Modeling Approach

Rady,E.A. and Zidan,A.I.

Abstract:

This paper proposes to illustrate the use of Dynamic Data System (DDS) approach to remove the deterministic trend and seasonality that causes nonstationary in sales as an economic time series. Modeling with three stage procedures deterministic, stochastic and combined model is used to decomposition nonstationary time series into two parts: one can be represented by (exponential and sinusoidal) as the deterministic functions depends on the time origin and another described by stochastic dynamical systems with autoregressive moving average ARMA(n,n-1) model. The results show a superiority for the combined model in reducing the mean square error (MSE) more than the traditional modeling approach uses seasonal difference operator which indicated that transformation model to stability.

Keywords:

Dynamic Data System (DDS), Deterministic trend, Periodic Trend, Seasonality, and Nonstationary Time Series.

1. **Introduction:**

In the classical approach of time series analysis, the procedure of simplifying the series of data by taking differencing or seasonality operators before modeling is often recommended in the literature. In such cases, modeling is based on identifying the model either from the data or from the plots of sample autocorrelation and partial autocorrelation. When trend and seasonality are dominated in the data, the sample autocorrelation fails to damp out quickly, and the plots of partial autocorrelation by the conventional methods are badly distorted, thus making it almost impossible to tentatively identify the model from their plots. The only way to get them to forms from which allow ordering autoregressive AR or moving average MA model can be guessed is to apply differencing or seasonality operators $(1-B^s)$ which in turn have to be guessed from the data, autocorrelations or partial autocorrelation. The danger of such indiscriminate operating or smoothing of the data simply for the sake of making it easier to analyze has been pointed out by Slutsky as pioneer author.

Such an operation itself may introduce spurious trends and periods in the resultant series that are absent in the original data. The final fitted model, although statistically adequate and apparently parsimonious, may give a completely distorted picture of the structure of the original series Kapoor et al. (1981).

The stationary time series models were based on the assumptions that the first two moments namely the mean and the covariance are independent of the time origin. These assumptions imply that the mean is fixed or constant, so that it may be subtracted and the series assumed to have zero mean, and that the covariance at a given lag depends only on the lag Box et al. (2015).

Some author, notably Box and Jenkins use the word "nonstationary" or trends for discrete ARIMA models with one or more roots with absolute value one, e.g. random walk, integrated random walk, EWMA, etc. One use of the term "nonstationary" when the nature of a series of data appears to be dependent on time origin follows the terminology in the stochastic process, system analysis, and control theory. The models for such a series of data, therefore, need to include a function which depends on time origin Chatfield (2016).

We will show in Dynamic Data System (DDS) modeling approach that nonstationary trends and seasonal pattern (periodic trends) in the data can be modeled by "***relaxing the first assumption of zero or fixed mean***". The first one, representing the mean of the time series, accounts for nonstationary trend by a deterministic function, which depends on the time origin. The second one is a stochastic part with zero mean so that it can be modeled by the method of ARMA (n, n-1).

The models for such nonstationary time series data, therefore, need to include a function that depends on the time origin. In this case, we will show that many of nonstationary data can be modeled by explicitly including polynomial, exponential, or sinusoidal function, dependent on the time origin, to represent the mean of the series. Such nonstationary trends can also be modeled by first subjecting the series to transformations such as differencing, either simple or seasonal, aimed to reduce the series to stationary. The transformation stationary series is then modeled by an ARMA model Pandit (1983). We, however, avoid such transformation in our modeling by DDS approach and put special emphasis on the system aspect of data. The combined model procedure of modeling nonstationary time series is developed for different kinds of trends such that exponential and periodic or seasonality.

2. Preliminary Modeling by DDS approach

The application used the monthly data for sales time series for Lydia Pinkham company in the period from January 1954-1960 i.e. 78 monthly observations. We use 60 sample from January 1954 to December 1958 for modeling and 18 samples from January 1959 to June 1960 for forecasting by conditional expectation. Sales time series is generally nonstationary. Figure (1)

show the plot of sales time series which appears the seasonality or periodic trend dominant in the data, so it is necessary to transform sales data to stationary time series.

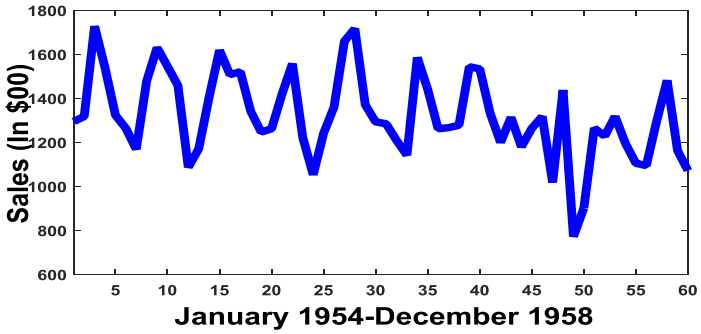


Figure (1): Sales

The propose of the preliminary stochastic modeling by DDS to detected the seasonality in data. A sequences of ARMA(n,n-1) apply on Monthly sales Lydia Pinkham data First, an ARMA(2,1) model is fitted by using a nonlinear least square routine, and then the order of the model is increased in steps of two, i.e., from (2,1) to (4,3) to (6,5) and so on. The statistical significance of the reduction in the sum of squares RSS after increasing the order of the model is checked by an F-test criterion, and the process of modeling was stopped when F-test become insignificant. The statistically adequate ARMA(n,n-1) models with all unified autocorrelation (#UAC) within the ± 2 possible band so the residuals a_t's can be approximately taken as independent and this should be further confirmed by F-test criteria to checked the reduction of RSS, a FORTRAN program was specially written for this purpose. The resulting is ARMA(6, 5) model. Table (1) shows the characteristics roots of the autoregressive operator with natural frequency and damping ratios.

Table (1): Characteristics Roots of the Autoregressive Operator

Discrete Complex Roots λ_i's	**Natural Frequency ω_n (HZ)**	**Damping Ratios ξ**	**Absolute Value of Roots**
-0.7096+/-0.3985	0.4199	0.0781	0.8138
0.5351+/-0.8436	**0.1600**	**0.0010**	**0.9990**
0.2144	0.2451	——	0.2144
0.9590	**0.0067**	——	**0.9590**

From the table absolute value of two λ_i's complex conjugate discrete roots (0.5351+/-0.8436) with lower natural frequency $\omega_n = \mathbf{0.1600}$ and high damping ratios $\xi = \mathbf{0.0010}$ have absolute value $0.999 \approx 1$ which lie on the unit circle indicating that the model is oscillatory unstable model contains periodic trend or seasonality dominant in the data. Then we apply the three stages modeling for nonstationary sales time series used the combined model to overcome the unit root problem.

3. Modeling for Nonstationary Time Series

The periodic trend or seasonality terminology used when the data shows the periodic tendency repeats regularly with time, which indicator that the data is nonstationary. The exponential trend with real exponent described the deterministic trend. If the exponential trend exponent is complex conjugate then, the nonstationary data has periodic trend called seasonality. In this case, another form of the combined model can be used for modeling purpose, which involved the (sin-cos) wave with specific amplitude and phase to express the periodic tendency which added to exponential trend part with pairs of the complex conjugate exponent.

The pairs of the complex conjugate of the exponential trend with imaginary parts of the roots represented the dominant frequency and it multiplies. So, the corresponding formulation of combined model with real exponent of exponential trend product of sin function with known or unknown period and phase angle recommended according to Feng and Sun(1988), Pandit et al.(1983),Kapoor et al.(1981). This process is mostly used in the field of linear system analysis, and this equivalent to treatment with nonstationary stochastic process analysis in the frequency domain where DDS modeling has been dubbed as " generalized Laplace transform" Pandit (1991).The combined model can be estimated using the initial values from estimation process for two parts deterministic and stochastic, then modeling procedures of nonstationary time series with deterministic trends and seasonality can be carried out in three stages as the following:

- **Stage1: Deterministic part**

A regular periodic tendency appears in the data refer to the fluctuating of the peak between months along of the year which focuses in some months and appears small in another, then the deterministic part of the model represented by exponential trend with real exponents for the growth trend and complex conjugate exponent for the periodic trend. The combined model in this case given by:

$$y_t = \sum_{j=1}^{\ell} R_j e^{r_j t} + \sum_{j=1}^{i} B_j e^{b_j t}[c_j \sin(j\omega t) + \sqrt{1-c_j^{\,2}}\cos(j\omega t)] + X_t \qquad \textbf{(1)}$$

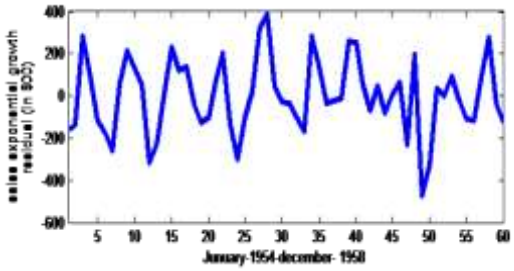
Where ℓ is the number of real exponents corresponding to growth trends, i is the number of pairs of complex conjugate roots corresponding to periodic trends, and $\ell + 2i = s$. R_j , r_j , B_j , b_j , are unknown parameters to be estimated, and ω is dominant frequency in radians per unit t . Both of R_j , r_j are exponential growth parameters and B_j , ψ_j denote to the amplitude and phase of periodic trend. The term of $e^{b_j t}$ the growth trend of periodic term and its harmonics, $c_j = \cos\psi_j$. X_t represent by the stochastic ARMA(n,n-1) model.

(1) Exponential Growth Trend

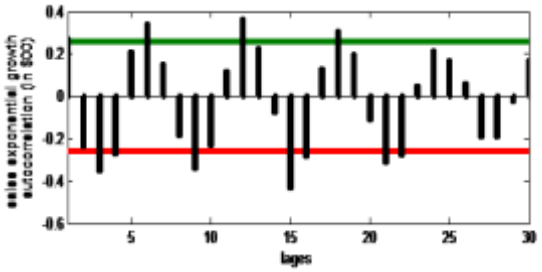
Estimation the combined model in equation (1) implementing through some steps: First fitting the exponential growth trend and then add the periodic trend, one by one according to the following model term:

$$y_t = R_1 e^{r_1 t} + \varepsilon_t \qquad \textbf{(2)}$$

The nonlinear least squares routine minimizing the sum of squares of ε_t's in equation (2). The plot of the residuals of the model equation (2) are shown in figure (2), the figure shows that the exponential growth represented by the actual data has been removed and the residuals only have periodic trends, but it still has out limits autocorrelation.



A. Sales Residual of Exponential growth



B. Sales Exponential growth autocorrelation

Figure (2): Residual and Autocorrelation of Sales Exponential Growth

(2) Addition of Periodic Trends

From the combined model in equation (1) the periodic trends can be added sequences one by one.

$$y_t = \sum_{j=1}^{\ell} R_j e^{r_j t} + \sum_{j=1}^{i} B_j e^{b_j t}\left[c_j \sin(j\omega t) + \sqrt{1-c_j^2}\cos(j\omega t)\right] + \varepsilon_t \qquad \textbf{(3)}$$

The criteria to choose the adequate model for periodic trend can be checked by continuing in modeling procedure by taking $\ell = 1$ and $i = 1,2,3,\ldots$ until the reduction in residual sum of squares RSS significant small and F- test show the adequacy of model. During increase i period, the estimated parameters obtained from each period inters as the initial for the next period. In our application, the residual sum of squares is reduced from **1849259.6** to **1768176.21** and the variance from 30820.99 to 29469.6. Since the improvement in the sum of squares is large, we successively fit the models for $\ell = 1$ and $i = 1,2,3,\ldots$ with $t = 6,4,3,\ldots.$,the results show in table (2) and adding to the combined model as show in column 3,4 and 5 in table (3).

(i)	Parameters	Deterministic Part **With** Exponential Growth Trend	Deterministic Part **Without** Exponential Growth Trend
(1)	$B_1 = -13.71$ $b_1 = 0.0326$ $C_1 = -0.997$ $\omega = \frac{2\pi}{12}$	$\lvert -13.71\rvert e^{0.0326t}(-0.999\sin\omega t + \sqrt{1-0.81}\cos\omega t)$ Figur: D.1	$(\sin\omega t + 175.56°)$ Figur: D.2
(2)	$B_2 = -222.31$ $b_2 = -0.0077$ $C_2 = 0.324$ $\omega = \frac{2\pi}{6}$	$\lvert -13.71\rvert e^{0.0326t}(-0.999\sin\omega t + \sqrt{1-0.81}\cos\omega t)$ Figur: D.3	$(\sin 2\omega t + 71.09°)$ Figur: D.4

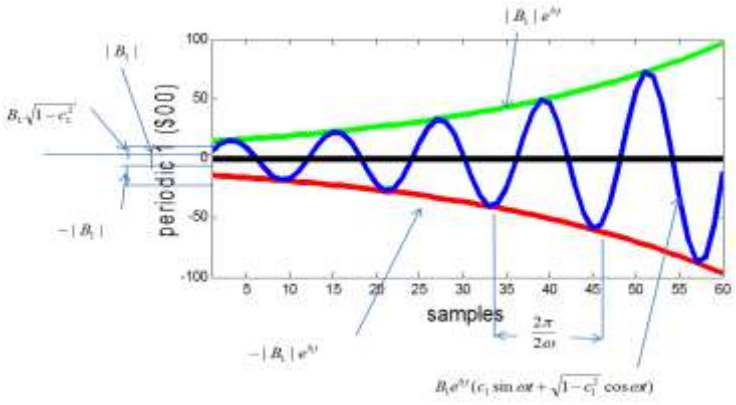


Figure D1: Sales First Periodic Term with Exponential Growth Trend

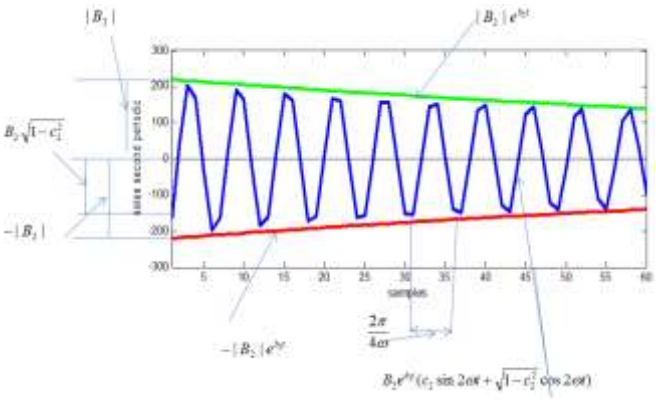


Figure D3: Sales Second Periodic Term With Exponential Growth

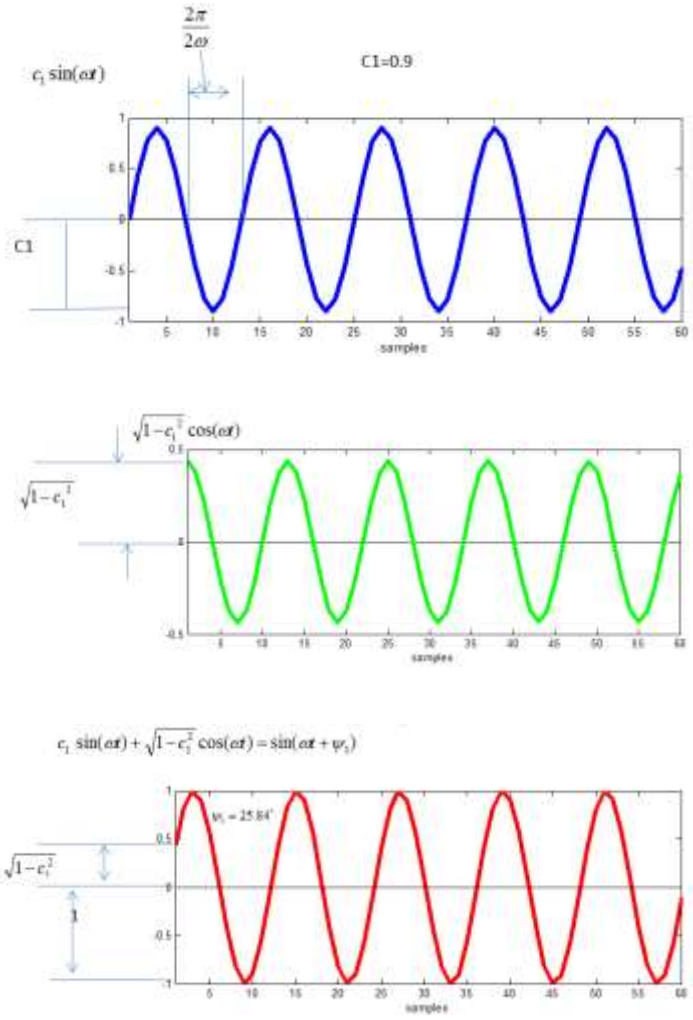


Figure D2: Sales First Periodic Term Without Exponential Growth

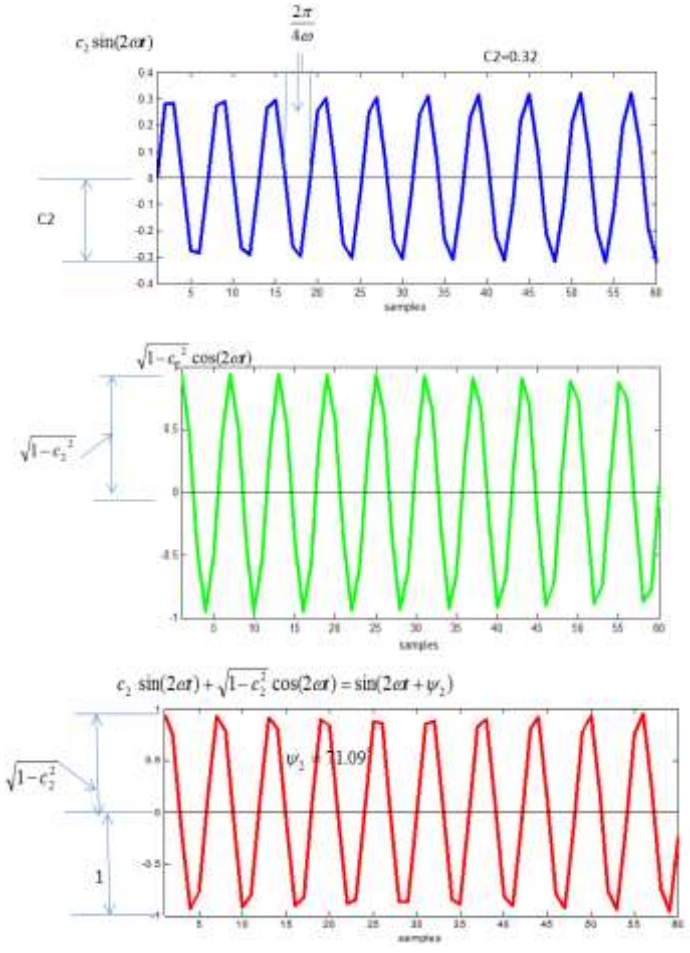


Figure D4: Sales Second Periodic Term Without Exponential Growth

Figure (3): Sales Periods for Deterministic Part

- **Stage2: Stochastic part**

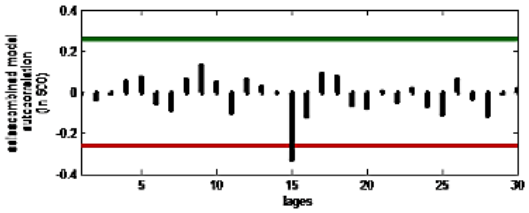
If the examination of the residual from the model in equation (3) shows independence, i.e. the residual become sequence of uncorrelated white noise refer to stationary time series. Thus, ARMA (n,n-1) can be represented the modeling of this residual as the stochastic part from decomposition process of the final combined model Feng and Sun(1988), Kapoor et al.(1981). The estimated parameters obtained from modeling stochastic part represented by ARMA (n,n-1) using as the initial values required to apply nonlinear least square NLS using in fitted the final combined model. The criteria to choose the adequate model for stochastic part is a significant reduction in residuals sum of squares obtained by continuant modeling procedure for sequence of models ARMA (2, 1), ARMA (4, 3), and ARMA (6, 5), and comparing two pairs of models to monitor the reduction of RSS using F-test significant .The result of our sales application can be shown in table (3) in column 6, 7, and 8.

- **Stage3: Combined Model: Deterministic Plus Stochastic**

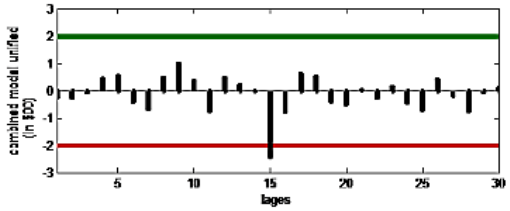
The formula of decomposition process of two stages parts of nonstationary time series which combined form deterministic part plus stochastic parts represented the adequate combined model with all parameters estimated can be written as:

$$\left.\begin{aligned} y_t &= \sum_{j=1}^{\ell=1} R_j e^{r_j t} + \sum_{j=1}^{i=2} B_j e^{b_j t}\left[c_j \sin(j\omega t) + \sqrt{1-c_j^{\,2}}\cos(j\omega t)\right] + X_t \\ X_t &= \sum_{s=1}^{4} \phi_s X_{t-s} - \sum_{w=1}^{3} \theta_w a_{t-w} + a_t \end{aligned}\right\} \quad \textbf{(4)}$$

Where: ϕ_s, θ_w are the parameters of the stochastic model and a_t become the residual of the combined model. The results of the estimation parameters for (2,4,3) the combined model are tabled in column 9 in table (3). The residuals of combined model reasonable white noise to which indicated that the combined model is equated.



(a) Combined Model Autocorrelation



(b) Combined Model Unified Autocorrelation

Figure (3): Sales Combined Model Autocorrelation and Unified Autocorrelation

Table (3): Seasonal Sales Combined Model Parameter

l: number of period ℓ : number of Exponentials (n, m) are order of ARMA model (l, n, m)	Deterministic (0,0,0)	Deterministic (1,0,0)	Deterministic (2,0,0)	Deterministic (3,0,0)	Stochastic (0,2,1)	Stochastic (0,4,3)	Stochastic (0,6,5)	Combined model (2,4,3)
Total numbers of parameters $=(2\ell+3l+m+n)$	2	5	8	13		7		15
Column number (1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
R_1	1462.51062 +/- 96.40159	1455.52551 +/- 99.95284	1461.71399 +/- 73.94735	1466.06079 +/- 78.41306				1458.18274 +/-75.90109
r_1	-0.00337 +/-0.00198	-0.00324 +/- 0.00210	-0.00343 +/- 0.00154	-0.00356 +/- 0.00167				-0.00341 +/- 0.00161
B_1		-34.35890 +/- 115.59333	-15.53420 +/- 60.17482	-9.63773 +/- 43.37028				-13.71927 +/- 59.78935
b_1		0.00839 +/- 0.08693	0.02823 +/- 0.08410	0.04188 +/_ 0.09115				0.03261 +/- 0.09176
c_1		-1.00001 +/- 0.08385	-1.00000 +/- 0.08209	-1.00016 +/- 0.19488				-0.99798 +/- 0.29114
B_2			-215.31850 +/_109.15339	-230.01367 +/- 119.85389				-222.31868 +/- 62.66507
b_2			-0.00791 +/_0.01665	-0.00969 +/- 0.01807				-0.00779 +/- 0.01353
c_2			0.30060 +/_0.47971	0.33313 +/- 0.48443				0.32473 +/- 0.21522
B_3				0.53716 +/- 11.79603				
b_3				0.07982 +/- 0.40749				
c_3				0.62403 +/- 15.62753				
ϕ_1					0.05993 +/- 10.193	0.26504 +/- 0.53128	-1.29664 +/- 0.70256	0.22332 +/- 0.56693
ϕ_2					0.02598 +/- 0.2717	-0.05775 +/- 0.59392	-0.48675 +/- 0.63312	0.01595 +/- 0.58224
ϕ_3						-0.34945 +/_0.53167	-0.20132 +/0.50006	-0.34428 +/- 0.53615
ϕ_4						-0.11905 +/_0.32770	-0.77785 +/_ 0.52249	-0.13240 +/- 0.37883
ϕ_5							-0.44192 +/_ 0.60910	
ϕ_6							-0.13521 +/_0.40987	
Θ_1					0.06345 +/-10.201	0.33846 +/-0.47221	-1.32087 +/- 0.64921	0.46583 +/- 0.68617
Θ_2						-0.12196 +/-0.52067	-0.51640 +/-0.48342	-0.23163 +/-0.65972
Θ_3						-0.80501 +/-0.47659	-0.65711 +/- 0.16006	-1.05445 +/-0.65325
Θ_4							1.49797 +/- 0.41146	
Θ_5							-0.74446 +/- 0.74008	
RSS	1894259.61	1768176.21	861432.67	901673.54	863458.43	675691.33	612910.57	491753.05
variance	30820.99	29469.60	14357.2	15026.22 Note: it is increased than (l=2) So we return to (l=2) as adequate period.	14390.97	11261.52	10215.17	8195.88
F-test for Stochastic model program calculate / standard Tabulate					-------	F(53,4) =3.682/2.61	F(49,4) =1.255/2.61 Calculate less than tabulate so we return to model (4,3)	

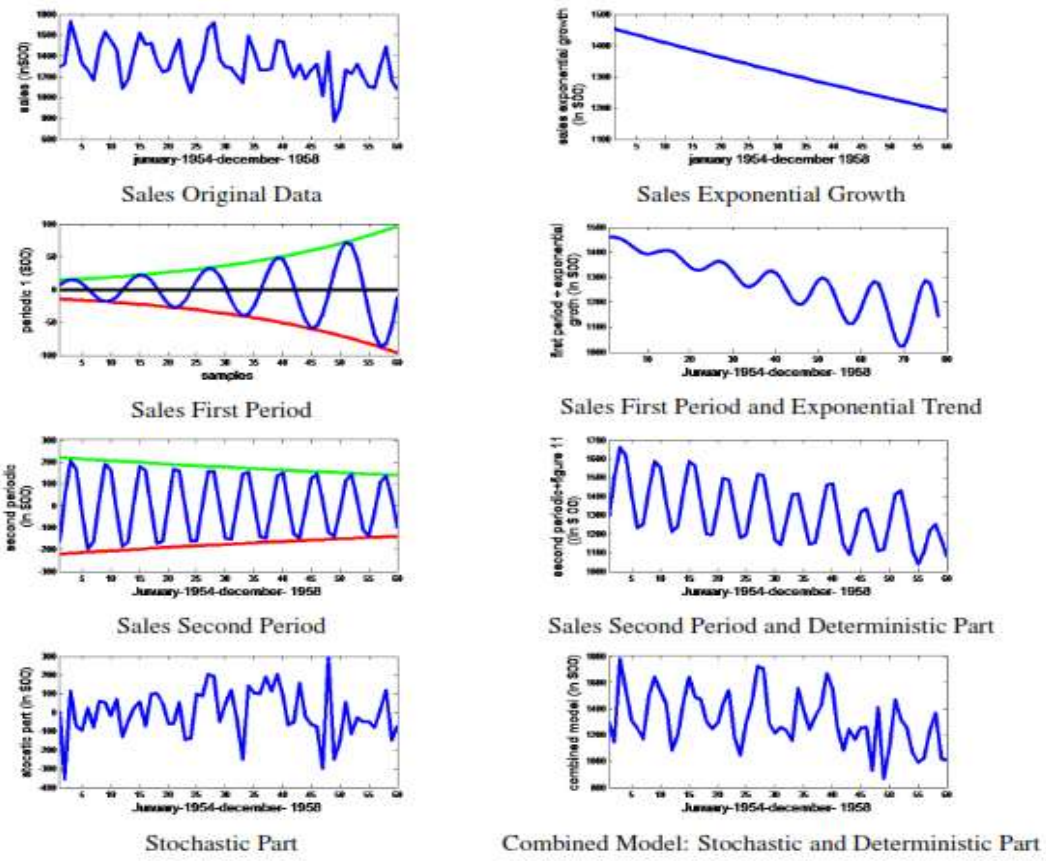


Figure (5): Three Stages Modeling for Seasonal Sales Combined Model

Table (4): Comparison of Goodness of Fit of Published Model and DDS approach for Lydia Pinkham Monthly Data (1954-1960)

Research Methodology	**Response Model**	**MSE Model Goodness of Fit**	**Rank**
Hanssens(1980) using **Univariate** ARIMA - **Sales**	$\left(1-B^{12}\right)Y_t = -44.98\left(1-0.257B^{12}-0.621B^{15}\right)a_t$	**18063.33**	**3**
Bhattacharya(1982) using **Univariate** ARIMA - **Sales**	$\left(1-0.4135B^{12}\right)(Y_t + 389.3511D_t) =$ $756.2111+\left(1+0.3615B-0.2485B^{2}-0.0353B^{3}\right)a_t$	**16384**	**2**
Rady,E.A and Zidan.A.I (2018) using **DDS** approach **univariate Sales** (2,4,3)Combined model	$y_t = \sum_{j=1}^{\ell=1} R_j e^{rjt} + \sum_{j=1}^{i=2} B_j e^{bjt}[c_j \sin(j\omega t)+\sqrt{1-c_j^{2}}\cos(j\omega t)+X_t$ $X_t = \sum_{s=1}^{4}\phi_s X_{t-s} - \sum_{w=1}^{3}\theta_w a_{t-w} + a_t$	**8195.88**	**1**

4. Conclusion:

A deterministic function depend on time origin integrated with stochastic dynamic model is used to modeling the nonstationary time series which the deterministic trend and seasonality are dominant indicated the nonstationary nature. Monthly sales data for Lydia Pinkham Company are analyzed to demonstrate the applicability of the modeling procedure for the combined model use Dynamic Data Systems (DDS) approach. Sales time series data appearing frequency peaks repeated at regular intervals describe the nonstationary components or seasonality. The deterministic component is approximated by (exponential and sinusoidal) deterministic function accounting for each period with specific amplitude and phase and the stochastic component is represented by ARMA(n,n-1) model. The new modeling with three stages procedure for combined model based on two principles, first, it avoids the trial and error of identifying stage to determine the orders of ARMA(n,m) model using Box-Jenkins approach which examination based on (ACF) autocorrelation and partial autocorrelation (PACF). Second, by using (DDS) we don't need to take the difference or any transform before modeling as recommended in Box-Jenkins approach which use nonstationary term for discrete ARMA with one or more roots absolute value one causes spurious trends and periods. In contrast DDS approach use the term of "nonstationary" when the nature of time series of data appears of data therefore need to include function which depend on the time origin. The comparisons results of mean square error (MSE) obtained from (2,4,3) combined model equal **8195.88** versus the difference approach modeling in previous literature in table (4) show that attainment high and accurate reduction in goodness of fit of (MSE) than Hanssens(1980) and Bhattacharya(1982) works equal **18063.33** and **16384** respectively.

References:

1. Bhattacharyya, M. N. (1982). Lydia Pinkham data remodelled. *Journal of time series analysis*, *3*(2), 81-102.
2. Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). Time series analysis: forecasting and control. John Wiley & Sons.
3. Chatfield, C. (2016). The analysis of time series: an introduction. CRC press.
4. Feng, X. and Sun, G. (1988). A new on-line approach for aids modeling and prediction through dynamic data systems identification (ddsi) method. In Engineering in Medicine and Biology Society, 1988. Proceedings of the Annual International Conference of the IEEE, pages 1084–1085. IEEE.
5. Hanssens, D. M. (1980). Bivariate time-series analysis of the relationship between advertising and sales. *Applied Economics*, *12*(3), 329-339
6. Kapoor, S. G., Madhok, P., and Wu, S. (1981). Modeling and forecasting sales data by time series analysis. Journal of Marketing Research, pages 94–100.
7. Pandit, S. M. (1991). Modal and spectrum analysis: data dependent systems in state space. Wiley-Interscience.
8. Pandit, S. M., Wu, S.-M., et al. (1983). Time series and system analysis with applications. Wiley New York

A Multivariate approach: Modeling and Forecasting with Economic Time Series Application

Rady,E.A. and Zidan,A.I.

Abstract:

This paper illustrates an application of a recently developed deterministic and stochastic modeling and time series analysis methodology called Dynamic Data Systems (DDS) to advertising-sales system forecasting. Univariate as well as simplified vector models called extended autoregressive moving average (EARMA) models are obtained for advertising-sales Lydia Pinkham Company. The conditional expectation of the statistically an adequate ARMA and EARMA models provides an accurate forecast for the peak value of sales. The usefulness of advertising as leading indicators is explored. The results show high improvement due to sales using advertising leading indicators.

Keywords:

Multivariate Time Series Analysis, Dynamic Data Systems (DDS), Extension autoregressive moving average (EARMA), Advertising-Sales system, conditional expectation forecast, and leading indicators forecast..

1. Introduction:

Multivariate time series analysis is the study of statistical models and methods of analysis that describe the relationships among several time series. For many time series arising in practice, a more effective analysis may be obtained by considering individual series as components of a vector time series and analyzing the series jointly. Such multivariate processes arise when several related time series is observed simultaneously over time, instead of observing just a single series as is the case in univariate time series analysis. Multivariate time series analysis is used when one wants to model and explain the interactions and co-movements among a group of time series variables (Economic indicators). Multivariate methods are very important in economics and much less so in other applications of forecasting. The multivariate view is central in economics, where single variables are traditionally viewed in the context of relationship to other variables.

Modeling multivariate time series data, which consist of simultaneous observations on several related variables of interest. This type of data commonly occurs in economic or business contexts where it may be thought that certain variables interact. The objective is usually to explore the dynamic relationships between the variables, perhaps to forecast the system by using a suitable representation. It can be difficult to build univariate models, so when we come to allow for between series effects in a suitable multivariate model we will meet even more problems. However, it is hoped that the extra information used may lead to a more representative model than a set of univariate models. Statistical Multivariate Time Series modeling methods include the vector autoregressive moving average process.

A Multivariate time series has more than one time-dependent variable. Each variable depends not only on its past values but also has some dependency on other variables. This dependency is used for forecasting future values. In forecasting and even in economics, multivariate models are convenient in modeling interesting interdependencies and achieve a better fit within a given data or economic indicator. Multivariate forecasting methods rely on models in the statistical sense of the word, though there have been some attempts at generalizing extrapolation methods to the multivariate case. This does not necessarily imply that these methods rely on models with regards to whether they are a theoretical, such as time series models. Much research has gone into the development of ways of analysis multivariate time series (MTS) data in both the modeling and forecasting, for more details see (Nasiri et al., 2017), (Onwukwe et al., 2014), (Sagaert et al., 2018), (Tsolacos et al., 2014), (Beracha et al., 2013).

Dynamic Data Systems (DDS) approach considering with a univariate time series as a realization of a stationary stochastic system. Such data can always be represented by a model of ARMA(n,n-1) form. The ARMA(n,n-1) models can be extended to vector autoregressive moving average model called EARMA(n,n,n-1) model to show that two or more sets of time series data, treated as the realization of vector stationary stochastic system, can be similarly represented by vector models (Pandit et al., 1983).

The most prominent and important application involving multiple sets of data arise is economic forecasting with leading indicators. We will be considered with models for discrete systems and use conditional expectation strategies for optimal forecasting in sense of reduction mean square error (MSE). Using leading indicators for business forecasting has been relatively rare, partly because our traditional time series methods do not readily allow incorporation of external variables. These indicators study time series fluctuations in different periods and are used to project the future status of the economy. We can define a leading indicator as a numerical variable that contains predictive information for our target variable (e.g., sales) at least as many periods in advance as the forecast lead time.

This paper is considering to applying the DDS multivariate approach to obtaining the EARMA model as the design for a single input-single output (SISO) system. A simpler form can be deduced from the general ARMAV model by assuming that the input and output noises are uncorrelated. This form of the vector autoregressive moving average model can greatly simplify the estimation of the parameters of input-output systems. The input and output noises in a bivariate autoregressive moving average model can be made uncorrelated by making. By transforming both the nonstationary input and output time series to stationary stochastic process with uncorrelated white noise without taking difference operator or any transform on the data to void the spurious regression that effecting in the forecast ability causes distortion. Then the EARMA model can be constructed from stationary time series and used for making leading indicators forecasting to improve the forecasting ability in the case of the univariate time series.

The multivariate analysis of the time series, which requires the application of several successive stages as follows: The first step, involved the modeling of the nonstationary time series with deterministic trend or seasonality by decomposition it into two parts, a stochastic part which describes autoregressive moving average dynamic ARMA(n,n-1) models and the other deterministic part which contained the exponential and sinusoidal functions represents the periodic frequency components with specific amplitudes and phases defined for each period to eliminate the periodic trend or seasonality and the deterministic trend. The second step, after converting the nonstationary time series by removing the periodic components and the deterministic trends into stationary time series, we obtain the head steps forecasting by using Conditional Expectation. The third step, use stationary time series obtained from the second step to modeling EARMA model as a special case of ARMAV(n,n-1) in bivariate case represent single-input single-output SISO system design. Finally, improve the system forecasting performance by leading indicators. The application implemented for economic advertising-sales system for Lydia Pinkham Company.

2. Modeling of Extended Autoregressive Moving Average EARMA model

ARMA models can be extended to multiple series. Assuming a two series model with one series representing an input X_{1t} and other an output X_{2t} of a system. The transfer function model which represents the relation between the input X_{1t} and error term a_{2t} and its effects on the output X_{2t} is taking the formulation**:**

$$X_{2t} = \frac{\phi_{211}B + \phi_{212}B^2 + \ldots + \phi_{21n}B^n}{1 - \phi_{221}B - \phi_{222}B^2 - \ldots - \phi_{22n}B^n} X_{1t} + \frac{1 - \theta_{221}B + \theta_{222}B^2 + \ldots + \theta_{22n}B^n}{1 - \phi_{221}B - \phi_{222}B^2 - \ldots - \phi_{22n}B^n} a_{2t} \quad \textbf{(1)}$$

This model called extended autoregressive moving average model denoted by EARMA(n,n,n-1), this means that there are two autoregressive variables of order n, and one moving average of order (n-1) describe the system dynamically. The general model can be written in matrix form as:

$$\begin{bmatrix} X_{1t} \\ X_{2t} \\ \vdots \\ X_{nt} \end{bmatrix} = \begin{bmatrix} \phi_{111} & \phi_{121} & \cdots & \phi_{1n1} \\ \phi_{211} & \phi_{221} & \cdots & \phi_{2n1} \\ \vdots & \vdots & \cdots & \vdots \\ \phi_{n11} & \phi_{n22} & \cdots & \phi_{nn1} \end{bmatrix} \begin{bmatrix} X_{1t-1} \\ X_{2t-1} \\ \vdots \\ X_{nt-1} \end{bmatrix} + \begin{bmatrix} \phi_{112} & \phi_{122} & \cdots & \phi_{1n2} \\ \phi_{212} & \phi_{222} & \cdots & \phi_{2n2} \\ \vdots & \vdots & \cdots & \vdots \\ \phi_{n11} & \phi_{n22} & \cdots & \phi_{nn1} \end{bmatrix} \begin{bmatrix} X_{1t-2} \\ X_{2t-2} \\ \vdots \\ X_{nt-2} \end{bmatrix} + \cdots + \begin{bmatrix} \phi_{11n} & \phi_{12n} & \cdots & \phi_{1nn} \\ \phi_{21n} & \phi_{22n} & \cdots & \phi_{2nn} \\ \vdots & \vdots & \cdots & \vdots \\ \phi_{n1n} & \phi_{n2n} & \cdots & \phi_{nnn} \end{bmatrix} \begin{bmatrix} X_{1t-n} \\ X_{2t-n} \\ \vdots \\ X_{nt-n} \end{bmatrix}$$

$$+ \begin{bmatrix} a_{1t} \\ a_{2t} \\ \vdots \\ a_{mt} \end{bmatrix} - \begin{bmatrix} \theta_{111} & \theta_{121} & \cdots & \theta_{1m1} \\ \theta_{211} & \theta_{221} & \cdots & \theta_{2m1} \\ \vdots & \vdots & \cdots & \vdots \\ \theta_{m11} & \theta_{m22} & \cdots & \theta_{mm1} \end{bmatrix} \begin{bmatrix} a_{1t-1} \\ a_{2t-1} \\ \vdots \\ a_{nt-1} \end{bmatrix} - \cdots - \begin{bmatrix} \theta_{11m-1} & \theta_{12m-1} & \cdots & \theta_{1mm-1} \\ \theta_{21m-1} & \theta_{22m-1} & \cdots & \theta_{2mm-1} \\ \vdots & \vdots & \cdots & \vdots \\ \theta_{m1m} & \theta_{m2m} & \cdots & \theta_{mmm-1} \end{bmatrix} \begin{bmatrix} a_{1t-m+1} \\ a_{2t-m+1} \\ \vdots \\ a_{mt-m+1} \end{bmatrix} \quad \textbf{(2)}$$

We can write in matrix notation as:

$$\mathbf{X_t = \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + \ldots + \Phi_n X_{t-n} + a_t - \Theta_1 a_{t-1} - \Theta_2 a_{t-1} - \ldots - \Theta_m a_{t-m}}$$

The essentials usefully for ERMA models as a special case of ARMAV are easily converted to state space models Olson et al. (1999), (Pandit et al., 1991).

The EARMA model as a single input-single output SISO system can be modeling by the same methodology for univariate ARMA(n,n-1) models with consideration by calculate the autocorrelation and cross-correlation of residuals between series which become the essential criteria in the case of multivariate series design by EARMA model to checking the basic assumption of independence of the residuals a_{jt} and $i \neq j$ which $\hat{\rho}_{ijk}$ lie within the $\pm \frac{2}{\sqrt{N}}$ band or the unified autocorrelation are less than two in magnitude (#UAC < 2), using the following relations:

$$\hat{\gamma}_{a_{ijk}} = \frac{1}{N} \sum_{t=k+1}^{N} a_{it} a_{jt-k} \quad , \quad \hat{\gamma}_{a_{ijk}} = \hat{\gamma}_{a_{ij(-k)}} \quad , \quad k \geq 0 \qquad \textbf{(3)}$$

$$\hat{\rho}_{ijk} = \frac{\hat{\gamma}_{a_{ijk}}}{\sqrt{\hat{\gamma}_{a_{ii0}} \hat{\gamma}_{a_{jj0}}}} \quad , \quad k = 0, \pm 1, \pm 2, \ldots \qquad \textbf{(4)}$$

The modeling procedure begin with obtain the initial value using inverse function method developed by Pandit, which provide the indicator of delay or dead time, and using nonlinear least square method to fit the models for input and output as a univariate series, and judge by the sum of square on increasing model order n, for more details see (Pandit et al., 1983). In many business and economic systems, one is often more interested in forecasting than control. However, the forecast of a series of interest can be improved by using information from a related series. Such as or, we can obtain an extended autoregressive moving average EARMA model for them related series is then called “leading indicator” treating the leading indicator series as or and the desired series by the procedure of modeling, then we can be used for forecasting or by the leading indicator.

3. Forecasting by A single Leading Indicator

In the case of single input- single output SISO, we assume the input or leading indicator or and the series as or. The relationship between the desired series and leading indicator can be represented by output or interest EARMA model which used to obtain the forecasting of the desired series or by or as a leading indicator. The general EARMA model which the feedback effects of interest series on the leading indicator required at least one lag, we can formulate the and model as:

$$\begin{aligned} X_{1t} &= \phi_{121}X_{2t-1} + \phi_{122}X_{2t-2} + \cdots + \phi_{12n}X_{2t-n} + \phi_{111}X_{1t-1} \\ &+ \phi_{112}X_{1t-2} + \cdots + \phi_{11n}X_{1t-n} - \theta_{111}a_{1t-1} - \theta_{111}a_{1t-2} \\ &- \cdots - \theta_{11m}a_{1t-m} + a_{1t} \end{aligned} \quad \textbf{(5)}$$

and, for the interest series or the output:

$$\begin{aligned} X_{2t} &= \phi_{21L}X_{1t-L} + \phi_{21L+1}X_{1t-L-1} + \cdots + \phi_{21n}X_{1t-n} + \phi_{221}X_{2t-1} \\ &+ \phi_{222}X_{2t-2} + \cdots + \phi_{222}X_{2t-n} - \theta_{221}a_{2t-1} - \theta_{221}a_{2t-2} \\ &- \cdots - \theta_{22m}a_{2t-m} + a_{2t} \quad , \quad 0 \le L \le n \end{aligned}$$

(6)

The forecast at lead times $\ell \le L$, the X_{1t} model is not needed since the forecast involve only the known present and past values of X_{1t}. Thus, the forecasting with the conditional expectation of X_{2t} can be written as:

$$\begin{aligned} \hat{X}_{2t}(\ell) &= \phi_{21L}X_{1t-L+\ell} + \phi_{21L+1}X_{1t-L-1+\ell} + \cdots + \phi_{21n}X_{1t-n+\ell} \\ &+ \phi_{221}\hat{X}_{2t}(\ell-1) + \phi_{222}\hat{X}_{2t}(\ell-2) + \cdots + \phi_{22\ell-1}\hat{X}_{2t}(1) \\ &+ \phi_{22\ell}X_{2t} + \phi_{22\ell+1}X_{2t-1} + \cdots + \phi_{22n}X_{2t-n} - \theta_{22\ell}a_{2t} - \theta_{22\ell}a_{2t-1} \\ &- \cdots - \theta_{22m}a_{2t-m} \quad , \quad \ell \le \ell \end{aligned} \quad \textbf{(7)}$$

From orthogonal decomposition, we can compute the forecasting errors and the variance of the step head forecast (Pandit et al., 1983), as the following:

$$e_{2t}(\ell) = \sum_{j=0}^{\ell-1} G_{22j}a_{2t+\ell-j} \quad \textbf{(8)}$$

Where G_{22j} denoted to the Green's function and the error variance is given by:

$$V[e_t(\ell)] = \gamma_{a22}\sum_{j=0}^{\ell-1} G_{22j}^2 \quad \textbf{(9)}$$

Then, the 95% probability limits on the forecasts of $x_{2t+\ell}$, $\ell \geq 2$ the interesting series x_{2t} is given by:

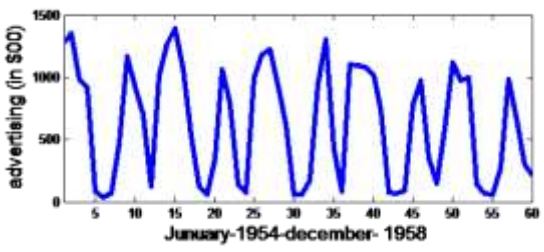
$$\hat{X}_{2t}(\ell) \pm 1.96[V(e_{2t}(\ell))]^{1/2} \quad \textbf{(10)}$$

That is:

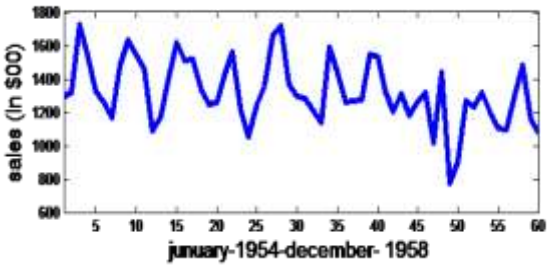
$$\hat{X}_{2t}(\ell) \pm 1.96[(1 + G_1^2 + G_2^2 + \cdots + G_{\ell-2}^2)(\phi_{211}^2 \gamma_{a11} + \gamma_{a22}) + (G_{\ell-1}^2 \gamma_{a22})]^{1/2} \quad \textbf{(11)}$$

4. Modeling Univariate Nonstationary Time Series

To design EARMA model as simplifying vector autoregressive moving average ARMAV from a nonstationary time series with deterministic trend and seasonality, it is necessary to transform each time series to stationary before construct EARMA model as recommended in literature (Rufino, 2008), (Wei et al., 2006). (Todorov et al., 2002), (Feichtinger et al., 1994), (Bhattacharyya, 1982). The advertising-sales system, the advertising time series denoted by representing the input of the system, and the sales as desired time series denoted by representing the output of the system. Lydia Pinkham data take from January 1954-June 1960, i.e. 78 monthly samples represent the total observations. We use 60 samples from January 1954-December 1958 for modeling, and 18 samples from January 1959-June 1960 for head steps forecasting. These time series are generally nonstationary and autocorrelated, figure (1) part a and b, shows the advertising and sales monthly data time series. The plot shows a regular periodic tendency which indicated that the seasonality is dominant in the data, so it is necessary to transform and from nonstationary or seasonality to stationary time series.



a. Advertising Time Series Lydia Pinkham Data



b. Sales Time Series Lydia Pinkham Data

Figure (1): Advertising-Sales Lydia Pinkham Data

Applying three stages procedures modeling, the deterministic part deepened on the time origin, the stochastic part represented by ARMA(n,n-1) model, and the combined model described the decomposition time series into two parts deterministic and stochastic in sense of reduce the residual sum of square (RSS) model, for nonstationary time series using DDS approach to obtain the combined model for each time series advertising and sales which transform to stationary see (Pandit et al., 1983). Then we obtain the results:

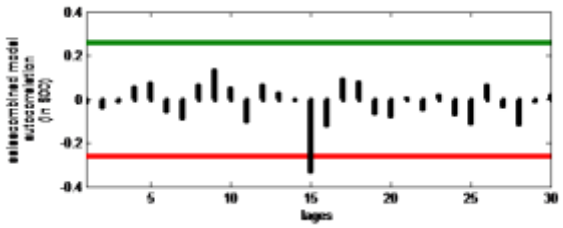
5. Sales Combined Model:

After the deterministic and stochastic part was estimated as table(1) column 3, checked by the autocorrelation which must not exceed about $(2/\sqrt{60})$ where 60 is the number of the samples and the unified autocorrelation not exceed about (± 2) as in figures (2) except at sample 15, which slightly exceed about the bound but still include in the 0:05% probability error band, so it can be considered the residuals reasonable white noise to this model which indicated that the combined model is equated, the sales combined model is given by:

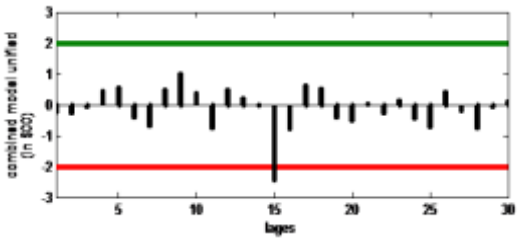
$$y_{2t} = \sum_{j=1}^{\ell=1} R_j e^{r_j t} + \sum_{j=1}^{i=2} B_j e^{b_j t}[c_j \sin(j\omega t + \sqrt{1-c_j^2}\cos(j\omega t)] + X_{2t}$$

$$x_{2t} = \sum_{s=1}^{4} \phi_s X_{1t-s} - \sum_{w=1}^{3} \theta_w a_{1t-w} + a_{2t} \qquad \textbf{(12)}$$

Where ℓ is the number of real exponents corresponding to growth trends, i is the number of pairs of complex conjugate roots corresponding to periodic trends, and $\ell + 2i = s$. R_j, r_j, B_j, b_j, are unknown parameters to be estimated, and ω is dominant frequency in radians per unit t, ϕ_s, θ_w are the parameters of the stochastic model and a_t become the residual of the combined model.



a) Combined model Autocorrelation



b) SalesCombined model Unified Autocorrelation

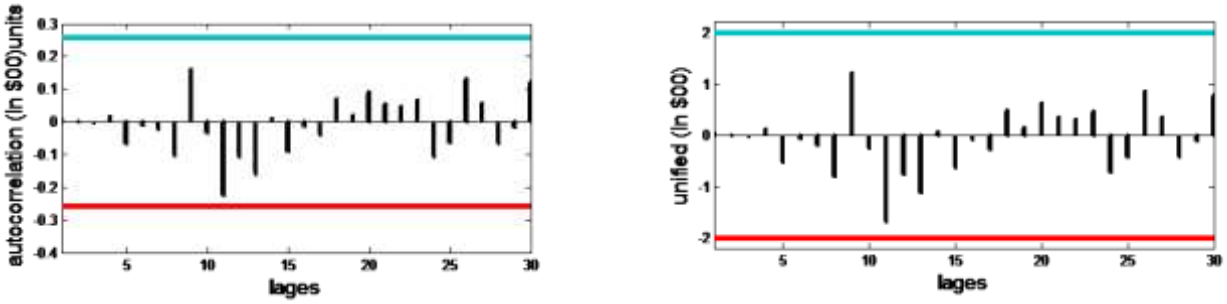
Figure (2): Sales Combined model Autocorrelation Unified Autocorrelation

6. Advertising Combined Model:

Similarly: the advertising combined model is given by:

$$y_{1t} = \sum_{j=1}^{\ell=1} R_j e^{r_j t} + \sum_{j=1}^{i=5} B_j e^{b_j t}[c_j \sin(j\omega t + \sqrt{1-c_j^2}\cos(j\omega t)] + X_{1t}$$

$$x_{1t} = \sum_{s=1}^{2} \phi_s X_{1t-s} - \sum_{w=1}^{1} \theta_w a_{1t-w} + a_{2t} \qquad \textbf{(13)}$$



a) Combined model Autocorrelation

b) Combined model Unified Autocorrelation

Figure (3): Advertising Combined model Autocorrelation Unified Autocorrelation

Table (1): Seasonal Sales Combined Model Parameter

Parameters	**Stage (1) Deterministic Part (2,0,0)**	**Stage(2) Stochastic Part (0,4,3)**	**Stage(3) Combined Model (2,4,3)**
R_1	1461.71399 +/- 73.94735		1458.18274 +/- 75.90109
r_1	-0.00343 +/- 0.00154		-0.00341 +/- 0.00161
B_1	-15.53420 +/- 60.17482		-13.71927 +/- 59.78935
b_1	0.02823 +/- 0.08410		0.03261 +/- 0.09176
c_1	-1.00000 +/- 0.08209		-0.99798 +/- 0.29114
B_2	-215.31850 +/- 109.15339		-222.31868 +/- 62.66507
b_2	-0.00791 +/- 0.01665		-0.00779 +/- 0.01353
c_2	0.30060 +/- 0.47971		0.32473 +/- 0.21522
ϕ_1		0.26504 +/- 0.53128	0.22332 +/-0.56693
ϕ_2		-0.05775 +/- 0.59392	0.01595 +/- 0.58224
ϕ_3		-0.34945 +/- 0.53167	-0.34428 +/- 0.53615
ϕ_4		-0.11905 +/- 0.32770	-0.13240 +/- 0.37883
Θ_1		0.33846 +/- 0.47221	0.46583 +/- 0.68617
Θ_2		-0.12196 +/- 0.52067	-0.23163 +/- 0.65972
Θ_3		-0.80501 +/- 0.47659	-1.05445 +/- 0.65325
RSS	**1768176.21**	**675691.33**	**491753.05**
MSE			**8195.88**

Table (2): Advertising Combined Model Parameter

Parameters	Stage (1) Deterministic Part (5,0,0)	Stage(2) Stochastic Part (0,2,1)	Stage(3) Combined Model (5,2,1)
R_1	716.58508 +/- 114.1733		716.22766 +/- 101.10614
r_1	-0.00435 +/- 0.0046		-0.00433 +/- 0.00423
B_1	368.22141 +/- 154.6523		367.56668 +/- 143.86305
b_1	-0.00576 +/- 0.0131		-0.00570 +/- 0.01238
c_1	0.54448 +/- 0.3953		0.54498 +/- 0.33838
B_2	-500.91730 +/- 155.9408		-501.60071 +/- 147.00570
b_2	-0.00328 +/- 0.0090		-0.00333 +/- 0.00863
c_2	-0.09953 +/- 0.3138		-0.09793 +/- 0.27823
B_3	108.78530 +/- 174.2372		107.17580 +/- 167.69550
b_3	-0.00842 +/- 0.0502		-0.00766 +/- 0.05078
c_3	0.99999 +/_ 0.0396		1.00000 +/- 0.03803
B_4	-220.47362 +/- 163.5739		-220.07288 +/- 155.73068
b_4	-0.00609 +/- 0.0224		-0.00608 +/- 0.02163
C_4	-0.75535 +/- 0.4861		-0.75258 +/- 0.47147
B_5	31.06175 +/- 171.016		29.50856 +/- 174.89449
b_5	-0.00848 +/- 0.1730		-0.00633 +/- 0.18745
C_5	-0.99967 +/- 0.2101		-1.00035 +/- 0.71068
ϕ_1		-0.29890 +/- 0.754	-0.71134 +/- 7.44574
ϕ_2		0.00161 +/- 1.2533	-0.03927 +/- 1.03649
Θ_1		-0.19150 +/- 0.7574	-0.61108 +/- 7.45602
RSS	**1330885.0**	**1306911.2**	1272495.5
MSE			**21208.2**

7. Sales Forecasting by Conditional Expectation

We interested in sales time series analysis for purpose of forecasting so computed the L steps ahead forecast from L = 1 to 18 monthly samples of sales Lydia Pinkham data. The forecasting consists of two parts, first the forecasting from the deterministic part without forecasting error which computed from the following model:

$$y_{2t} = \sum_{j=1}^{\ell=1} R_j e^{r_j t} + \sum_{j=1}^{i=2} B_j e^{b_j t}[c_j \sin(j\omega t + \sqrt{1-c_j^2}\cos(j\omega t)] + \varepsilon_t \quad \textbf{(14)}$$

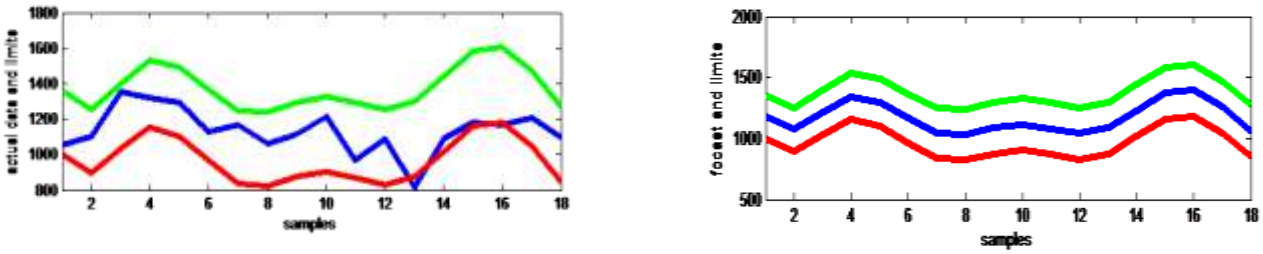
and the second part stochastic forecasting using conditional expectation which will be obtained from the stochastic part of the combined model as ARMA(n,n-1) given by:

$$x_{2t} = \sum_{s=1}^{4} \phi_s X_{1t-s} - \sum_{w=1}^{3} \theta_w a_{1t-w} + a_{2t} \quad \textbf{(15)}$$

The sales forecasting for 18 a head steps forecast with 95% probability limit show in the following table.

Table (3): Sales Forecasting by Conditional Expectation

Samples number	Actual values	Forecasted Values	Error =actual-forecast	95% Lower limits	95% Upper Limits
61	1052.000	1178.95740	-126.9574	1001.51868	1356.39612
62	1102.000	1074.34753	27.65247	894.96582	1253.72925
63	1355.000	1216.51050	138.4895	1034.41174	1398.60925
64	1323.000	1346.02441	-23.02441	1157.41602	1534.63281
65	1296.000	1298.67029	-2.670288	1103.39697	1493.94360
66	1127.000	1166.63013	-39.63013	966.02338	1367.23694
67	1170.000	1043.52087	126.4791	839.04779	1247.99390
68	1059.000	1030.90442	28.09558	823.75104	1238.05786
69	1116.000	1084.56909	31.43091	875.59674	1293.54150
70	1214.000	1114.43445	99.56555	904.24042	1324.62842
71	966.0000	1080.08252	-114.0825	869.07220	1291.09277
72	1089.000	1041.24841	47.75159	829.69397	1252.80286
73	814.0000	1087.06616	-273.0662	875.14948	1298.98279
74	1087.000	1229.44763	-142.4476	1017.28986	1441.60535
75	1180.000	1372.41870	-192.4187	1160.10059	1584.73682
76	1167.000	1393.27222	-226.2722	1180.84729	1605.69714
77	1210.000	1257.92859	-47.92859	1045.43262	1470.42456
78	1092.000	1056.46277	35.53723	843.91962	1269.00598
Sum of Squares error (RSS)			**267793.1**		
Mean Square Error MSE			**14877.394**		

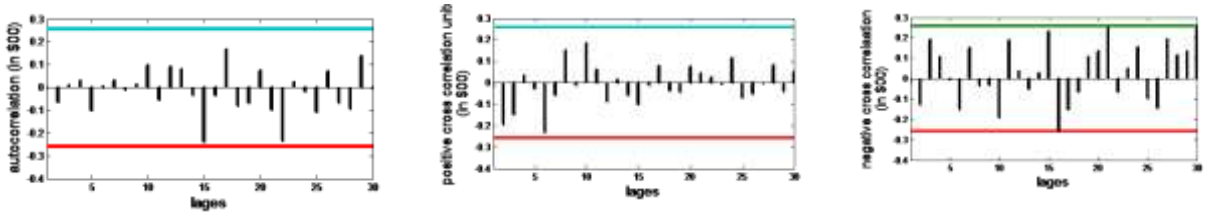


a) Actual Sales Data with Forecasting Limit b) Sales Forecasting With Limit

Figure(4): Univariate Sales Forecasting

8. Modeling Bivariate Advertising-Sales (SISO) System

After treatments with the univariate nonstationary advertising and sales time series and converting both from nonstationary to stationary time series with autocorrelation and unified autocorrelation lies in possible band as in figures (2) and (3) this procedure intended to obtain the adequate EARMA model for advertising X_{1t} as input or as leading and sales X_{2t} as output or as desired series. We used general procedures of modeling using DDS approach as illustrated in section (2) this design called single input single output (SISO). We choose the adequate (4,4,3) EARMA sales model and (1,1,0) for advertising in sense of minimum RSS, the autocorrelation as well as positive-negative lag cross-correlation lies within $2/\sqrt{60}$ possible band, or (the unified correlation is less than two in magnitude), and significant F-test is used as a criteria's to obtain the adequate EARMA model. Figure (5) shows the sales autocorrelation, positive lag cross-correlation, negative lag cross-correlation which appear that all of it lies within the possible limits.



a) Autocorrelation b) Positive lag cross-correlation c) Negative lag cross-correlation

Figure(5): Advertising-Sales EARMA Model

The system equations of bivariate adequate advertising-sales EARMA(n,n,n-1) model is (1;1; 0) for advertising and (4;4; 3) for sales given by:

$$\left.\begin{aligned} x_{1t} &= -0.132x_{1t-1} + 0.136x_{2t-1} + a_{1t} \\ x_{2t} &= 0.165x_{1-1t} + 0.321x_{1t-2}\ 0.062x_{1t-3} - 0.254x_{1t-4} \\ &\quad -0.10x_{2t-1} + 0.274x_{2t-2} + 0.058x_{2t-3} - 0.133x_{2t-4} \\ &\quad +0.22\ a_{2t-1} - 0.40a_{2t-2} + 0.7a2t\text{-}3 + a_{2t} \end{aligned}\right\} \quad \textbf{(16)}$$

9. ERMA Model Stability

We study the stability of the ERMA model in two directions. First, the roots of the model (4;4; 3), the roots and their natural frequencies with damping ratios for EARMA model are listed in table (4) which shows that all AR roots lie in unit circle because the absolute values for it less than one and equal 0.44, and 0.64 respectively which indicated that the system stability, see Olson et al.(1999).

Table (4): Sales EARMA Characteristic Roots of Autoregressive Operator

Discrete roots	**Natural Frequency** ω_n	**Damping Ratios** ξ	**Absolute value of roots**
0.4762+/- 0.3853	0.1334	0.5846	**0.6125**
-0.5270+/- 0.2775	0.4308	0.1914	**0.5956**

The second direction is the Green's function and impulse response function between the series from advertising to sales which take the form:

a) Impulse Response Function of Series 2 to series 1

$$IMP(J) = 0.53145*(0.61255)**J*COS\left[2*P1*0.10826*DELTA*J+(-2.50648)\right] + 0.62667*(0.59562)**J*COS\left[2*P1*0.42285*DELTA*J+(-0.33055)\right] \quad \text{... (17)}$$

b) **Green's Function of Series 2 with its residuals**

$$G(J) = 0.93959*(0.61255)**J*COS\left[2*P1*0.10826*DELTA*J+(1.95845)\right] + 2.76357*(0.59562)**J*COS\left[2*P1*0.42285*DELTA*J+(-1.05828)\right] \quad \text{...(18)}$$

The impulse response function of series 2 to series 1, and the Green's function of series 2 to its residuals appear in figure (6) are decayed to zero described the model stability.

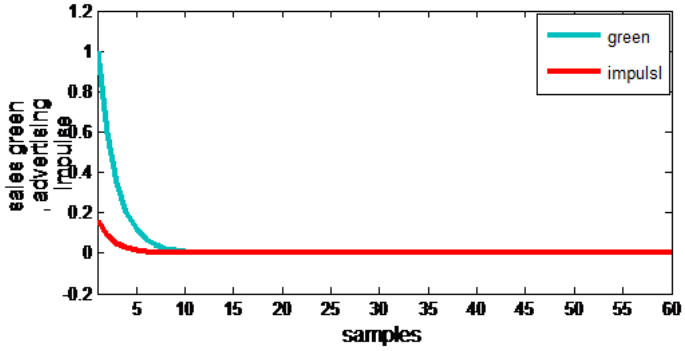


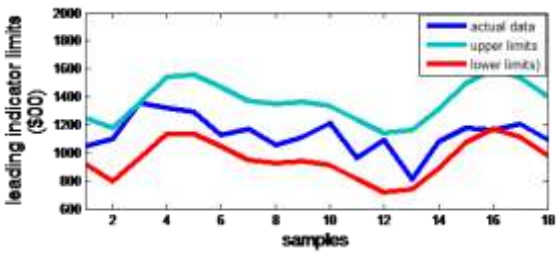
Figure (6): Sales Green's function and impulse from Advertising

10. Forecasting by Leading Indicator

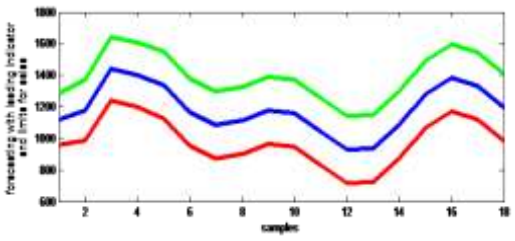
The conditional expectation forecasting for univariate time series was an illustration in previous sections (3). We obtain the bivariate stationary Lydia Pinkham advertising-sales system which the forecasting for system output or the sales time series as desired can be improved by using a related time series; such a related time series is called a ***leading indicator***, in current thesis application, we take the advertising is leading indicator to sales. The application design SISO system described the advertising time series denoted by x_{1t} as system input, and the sales time series denoted by x_{2t} as the system output. The equation (16) represents the EARMA model of the advertising-sales system. Computation of Conditional expectation for x_{1t}, x_{2t} according to the rules of conditional expectation is given by :

$$\left.\begin{aligned}\hat{x}_{1t}(\ell) &= -0.132\hat{x}_{1t}(\ell-1) + 0.136\hat{x}_{2t}(\ell-1) \\ \hat{x}_{2t}(\ell) &= 0.165\hat{x}_{1t}(\ell-1) + 0.321\hat{x}_{1t}(\ell-2) - 0.062\hat{x}_{1t}(\ell-3) - 0.254\hat{x}_{1t}(\ell-4) \\ &\quad -0.10\hat{x}_{2t}(\ell-1) + 0.274\hat{x}_{2t}(\ell-2) + 0.058\hat{x}_{2t}(\ell-3) - 0.133\hat{x}_{2t}(\ell-4)\end{aligned}\right\} \quad \textbf{(19)}$$

Then, the results of the sales forecasting by advertising leading indicator give in table (5).



1. Aactual data with leading indicator forecasting limits



2. leading indicator Forcasting with limits

Figure (7): Advertising-Sales leading indicator Forcasting

Table (5): Sales with Advertising leading Indicator

Samples number	Sales Actual values	Forecast Without leading	Error without Leading	Forecast With leading	Error with leading	95% Lower limits	95% Upper limits
61	1052.000	1178.95740	-126.9574	1121.53174	-69.53174	956.58356	1286.47986
62	1102.000	1074.34753	27.65247	1178.70557	-76.70557	987.22095	1370.19019
63	1355.000	1216.51050	138.4895	1439.58215	-84.58215	1239.70264	1639.46167
64	1323.000	1346.02441	-23.02441	1402.57385	-79.57385	1199.86194	1605.28577
65	1296.000	1298.67029	-2.670288	1336.71460	-40.71460	1125.34961	1548.07959
66	1127.000	1166.63013	-39.63013	1164.06616	-37.06616	952.36639	1375.76587
67	1170.000	1043.52087	126.4791	1085.06543	84.93457	873.25085	1296.88000
68	1059.000	1030.90442	28.09558	1115.96436	-56.96436	904.11035	1327.81836
69	1116.000	1084.56909	31.43091	1175.94971	-59.94971	964.08215	1387.81726
70	1214.000	1114.43445	99.56555	1161.54834	52.45166	949.67615	1373.42053
71	966.0000	1080.08252	-114.0825	1047.03235	-81.03235	835.15857	1258.90613
72	1089.000	1041.24841	47.75159	930.40320	158.5968	718.52887	1142.27747
73	814.0000	1087.06616	-273.0662	933.89221	-119.8922	722.01770	1145.76672
74	1087.000	1229.44763	-142.4476	1084.81409	2.185913	872.93951	1296.68860
75	1180.000	1372.41870	-192.4187	1282.52625	-102.5262	1070.65161	1494.40088
76	1167.000	1393.27222	-226.2722	1384.78076	-217.7808	1172.90613	1596.65540
77	1210.000	1257.92859	-47.92859	1333.20032	-123.2003	1121.32568	1545.07495
78	1092.000	1056.46277	35.53723	1195.77649	-103.7765	983.90192	1407.65112
Sum of Squares Error (**RSS**) for forecast			**267793.1**		**174026.2 Reduced by 35%**		
Mean Square Error (**MSE**)			**14877.394**		**9668.1222**		

By using the impulse response function $IMP(J)$ of series 2 to series 1 in equation (17), and $G(J)$ Green's function of Series 2 in equation (18). The upper and lower probability limits leading indicator forecasting error appear in last two are plots in the figure (7) which show that the forecast values are within in the possible limits. The results show that the forecasting of sales time series with advertising leading indicator is better than the forecasting for univariate sales time series by conditional expectation with 35% improve percentage.

Table (6): Comparison of Goodness of Fit AND Forecasting Ability of Published Model and DDS approach for Lydia Pinkham Monthly Data (1954-1960)

Research Methodology	**Response Model**	**MSE Model Goodness of Fit**	**Rank**	**Forecast Ability**	**Rank**
Hanssens(1980) using **Univariate** ARIMA - Sales	$\left(1-B^{12}\right)Y_t = -44.98\left(1-0.257B^{12}-0.621B^{15}\right)a_t$	18063.33	3	19876.22	2
Bhattacharya(1982) using **Univariate** ARIMA - Sales	$\left(1-0.4135B^{12}\right)(Y_t + 389.3511D_t) = 756.2111 + \left(1+0.3615B-0.2485B^2-0.0353B^3\right)a_t$	16384	2	27395.94	3
Rady,E.A. and Zidan.A.I (2018) using DDS approach **Univariate** Sales Combined model (2,4,3)	$y_t = \sum_{j=1}^{\ell=1} R_j e^{r_j t} + \sum_{j=1}^{i=2} B_j e^{b_j t}[c_j \sin(j\omega t)+\sqrt{1-c_j^2}\cos(j\omega t)+X_t$ $X_t = \sum_{s=1}^{4}\phi_s X_{t-s} - \sum_{w=1}^{3}\theta_w a_{t-w} + a_t$	8195.88	1	**14877**	1
Hanssens(1980) **Bivariate** Advertising-**Sales** Rational lag Distributed Structural (RSF) –	$\nabla^{12}S_t = \frac{0.374(0.126)B}{1-0.546B}\nabla^{12}A_t + \left(1+0.082B-0.068B^2+0.372B^3\right)e_{2t}$	25600	3	19698.72	2
Bhattacharya(1982) **Bivariate** Advertising-**Sales**	$(Y_t + 419.4913D_t - 211.7867L_t) = 1179.1776 + ((0.2329B)X_t + \left(1+0.1182B-0.1230B^2\right)\eta_{1,t}$	11881	2	41345	3
Rady,E.A. and Zidan.A.I (2018) **Bivariate** using DDS approach EARMA(1,1,0) Advertising –**Sales** (4,4,3)	$x_{1t} = -0.132x_{1t-1} + 0.136x_{2t-1} + a_{1t}$ $x_{2t} = 0.165x_{1-1t} + 0.321x_{1t-2}\ 0.062x_{1t-3} - 0.254x_{1t-4} - 0.10x_{2t-1} + 0.274x_{2t-2} + 0.058x_{2t-3} - 0.133x_{2t-4} + 0.22a_{2t-1} - 0.40a_{2t-2} + 0.7a2t-3 + a_{2t}$	**7084.8**	1	**9668**	1

▪ Conclusion

After converting the nonstationary time series by removing the periodic components and the deterministic trends into stationary time series, we obtained the head steps forecasting by using Conditional Expectation. The third step, use stationary time series obtained from the second step to modeling EARMA model as a special case of VARMA (n,n-1) in bivariate case represent single-input single-output SISO system design. The fourth step, improve the bivariate system performance by the leading indicator. The results also showed that the effectiveness of the frequency domain approach in the bivariate analysis of the nonstationary time series.

The continuous transform process, such as taking the differences as in the classic approach (Box-Jenkins) of time series analysis was avoided by decomposition the nonstationary time series with stochastic and deterministic components which described Frequency components in the form of sinusoidal function models as periodic functions with frequencies contributing to the removal of seasonal and deterministic trends. The results of advertising-sales economic system application using DDS approach showed the effectiveness of forecasting by using the leading indicators in improving the system performance with optimum values of the parameters which contribution in reducing the MSE by 35% than the conditional expectation forecasting. Robustness of DDS approach appear in both modeling and forecasting in univariate case of "nonstationary" when the nature of time series of data appears of data, therefore, need to include function which depends on the time origin as well as bivariate analysis with EARMA design SISO, the comparisons results of mean square error (MSE) obtained from (2,4,3) combined model versus the difference approach modeling in previous literature in table (5), and forecasting with conditional expectation show that attainment high and accurate reduction in goodness of fit and forecasting ability of (MSE) than (Hanssens, 1980) and (Bhattacharya, 1982) works.

Acknowledgements

The authors would like to thank the reviewers for their valuable comments on the manuscript.

Reference:

Beracha, E., & Wintoki, M. B. (2013). Forecasting residential real estate price changes from online search activity. *Journal of Real Estate Research, 35*(3), 283-312.

Bhattacharyya, M. N. (1982). Lydia Pinkham data remodeled. Journal of time series analysis, 3(2):81–102.

Feichtinger, G., Hartl, R. F., and Sethi, S. P. (1994). Dynamic optimal control models in advertising: recent developments. Management Science, 40(2):195–226.

Hanssens, D. M. (1980). Bivariate time-series analysis of the relationship between advertising and sales. *Applied Economics*, *12*(3), 329-339.

Nasiri, H., Taghizadeh, K., Amiri, B., & Shaghaghi Shahri, V. (2017). Developing Composite Leading Indicators to Forecast Industrial Business Cycles in Iran. *International Journal of Research in Industrial Engineering*, *6*(1), 69-89.

Olson, W. W., Filipovic, A., Sutherland, J., and Pandit, S. (1999). Reduction of the environmental impact of essential manufacturing processes. Technical report, SAE Technical

Paper.

Onwukwe, C. E., & Nwafor, G. O. (2014). A Multivariate Time Series Modeling of Major Economic Indicators in Nigeria. *American Journal of Applied Mathematics and Statistics*, *2*(6), 376-385.

Pandit, S. M. (1991). Modal and spectrum analysis: data dependent systems in state space. Wiley-Interscience.

Pandit, S. M. (1991). *Modal and spectrum analysis: data dependent systems in state space*. New York: Wiley.

Pandit, S. M., Wu, S.-M., et al. (1983). Time series and system analysis with applications. Wiley New York.

Rajurkar, K. and Nissen, J. (1985). Data-dependent systems approach to short-term load forecasting. IEEE transactions on systems, man, and cybernetics, (4):532–536.

Rufino, C. C. (2008). Lagged effect of tv advertising on sales of an intermittently advertised product. Browser Download This Paper.

Sagaert, Y. R., Aghezzaf, E. H., Kourentzes, N., & Desmet, B. (2018). Tactical sales forecasting using a very large set of macroeconomic indicators. *European Journal of Operational Research*, *264*(2), 558-569.

Todorov, E. and Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. Nature neuroscience, 5(11):1226.

Tsolacos, S., Brooks, C., & Nneji, O. (2014). On the predictive content of leading indicators: the case of US real estate markets. *Journal of Real Estate Research*, *36*(4), 541-573.

Wei, W. W. et al. (2006). Time series analysis: univariate and multivariate methods. Pearson Addison Wesley.

Statistical Inference of Geometric Distribution under Type Ⅰ Censoring Sample with Missing Data

Ahmed A. El-sheikh[1], Naglaa A. Mourad[1] , Alaa S. Shehataa[1]

Abstract

The parameters of two Geometric distribution populations are estimated and the hypothesis testing on the equality of parameters are constructed under type I censoring with missing data. A confidence interval is placed. The consistency and asymptotic normality of the estimators are proved.

Keywords: Asymptotic Normality, Consistency, Geometric distribution, Maximum likelihood, Missing data, Type Ⅰ censored sample.

Introduction

The problem of estimation of parameters with missing data is very common in many studies and field experiments. Some researchers estimated parameters for two populations for different distributions and tested the hypothesis on the equality of two parameters under type I censoring with missing data. Zhao et al, (2009) estimated parameters and tested hypothesis of means of two exponential populations under type I censoring sample when data are missing. A variety of methods have been developed to estimate the unknown parameters of different models when some data are missing. In this paper, we estimate the parameters of two Geometric distribution populations and test the hypothesis on the equality of the parameters under type I censoring in case of missing data.

This paper is organized as follows. In Section 1, the parameters of two geometric distribution populations are estimated. In Section 2, consistency and normality property for estimators are proved. The hypothesis testing and confidence interval of parameters in two populations are discussed in Section 3.

[1]Department of applied Statistics and Econometrics, Institute of statistical studies and research, Cairo University, Egypt

1. Estimation of Parameters in case of Geometric distribution under type I censoring with Missing Data

In this section, estimators of parameters are derived by using maximum likelihood method in case of two geometric distribution populations under type I censoring when some data are missing. The consistency and normality of the estimators are proved.

The probability density function of geometric distribution has the form:

$$\boldsymbol{P(X = x) = (1 - \theta)^{x-1}\theta, \quad x = 1, 2, \ldots.}$$

$$\boldsymbol{E(x) = \frac{1}{\theta}, \quad v(x) = \frac{1 - \theta}{\theta^2}}$$

where θ_i ,i =1,2 i are unknown parameters that presented two geometric populations for n independent observations. The first sample is denoted as $(x_1, x_2, \ldots, x_n)$ with parameter θ_1 (unknown). The actual observed data is $(z_i, \alpha_i, \delta_i)$ $(i = 1,2, \ldots, n)$, where $(z_1, z_2, \ldots, z_n)$, $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ and $(\delta_1, \delta_2, \ldots, \delta_n)$ are independent. Assume that $z_i = \min(T_0, x_i)$,

$\alpha_i = I\{x_i \leq T_0\} - I\{x_i > T_0\}$ be the censoring or status indicator for T_0

where T_0 be a predetermined time to terminate the experiment

$$\alpha_i = \begin{cases} 1 & x_i \leq T_0, z_i \text{ is observed} \\ -1 & x_i > T_0, z_i \text{ is censored} \end{cases}$$

and

$$\delta_i = \begin{cases} 1 & z_i \text{ is observed with } p(\delta_i = 1) = p \\ 0 & \text{otherwise} \end{cases}$$

Similarly, the second sample is denoted as $(y_1, y_2, \ldots, y_n)$ with parameter θ_2 (unknown). The actual observed data is (M_i, β_i, η_i) $(i = 1,2, \ldots, n)$, where $(M_1, M_2, \ldots, M_n)$, $(\beta_1, \beta_2, \ldots, \beta_n)$ and $(\eta_1, \eta_2, \ldots, \eta_n)$ are independent. Assume that $M_i = \min(T_0, y_i)$,

$\beta_i = I\{y_i \leq T_0\} - I\{y_i > T_0\}$ be the censoring or status indicator for T_0

where

$$\beta_i = \begin{cases} 1 & y_i \leq T_0, M_i \text{ is observed} \\ -1 & y_i > T_0, M_i \text{ is censored} \end{cases}$$

and

$$\eta_i = \begin{cases} 1 & M_i \text{ is observed with } p(\eta_i = 1) = p \\ 0 & \text{otherwise} \end{cases}$$

1.1Estimation of Parameters by Maximum Likelihood Method in Case of One Population

In this section, the maximum likelihood method will be used to estimate the parameters of Geometric distribution in case of one population when some data are missing.

The maximum likelihood function has the form:

The M.L function has the form:

$$L_1(\theta_1) = \prod_{i=1}^{n} (f(z_i;\theta_1))^{A_i} (S(z_i;\theta_1))^{B_i} \quad (1)$$

where $S(z_i;\theta_1) = 1 - F(z_i) = (1-\theta_1)^{z_i}$ be survival function,

$A_i = \alpha_i\delta_i(\alpha_i\delta_i + 1)/2$ and $B_i = \alpha_i\delta_i(\alpha_i\delta_i - 1)/2$, i=1,2,…,n

$$L_1(\theta_1) = \prod_{i=1}^{n} ((1-\theta_1)^{z_i-1}\theta_1)^{A_i} ((1-\theta_1)^{z_i})^{B_i} \quad (2)$$

Hence, the logarithm of the likelihood function is given by

$$\ln L_1(\theta_1) = \sum_{i=1}^{n} \left[A_i\left((z_i - 1)\ \ln(1-\theta_1) + \ln\theta_1\right) + B_i z_i \ln(1-\theta_1)\right] \quad (3)$$

$$\frac{\partial(\ln L_1(\theta_1))}{\partial\theta_1} = \sum_{i=1}^{n} \left[\left(\frac{A_i}{\theta_1} - \frac{A_i(z_i-1)}{1-\theta_1}\right) - \frac{B_i z_i}{1-\theta_1}\right] \quad (4)$$

$$\sum_{i=1}^{n} A_i - \theta_1\left(\sum_{i=1}^{n} A_i z_i + \sum_{i=1}^{n} B_i z_i\right) = 0$$

The estimator of θ_1 will be:

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n} A_i}{\sum_{i=1}^{n}(A_i + B_i)z_i} \quad (5)$$

In a similar way, The estimator of θ_2 will be:

$$\hat{\theta}_2 = \frac{\sum_{i=1}^{n} C_i}{\sum_{i=1}^{n}(C_i + D_i)M_i} \quad (6)$$

$C_i = \beta_i\eta_i(\beta_i\eta_i + 1)/2, and\ D_i = \beta_i\eta_i(\beta_i\eta_i - 1)/2$, i=1,2,…,n

1.2 Estimation of Parameters by Maximum Likelihood Method in Case of Two Populations

Assume $H_0 = \theta_1 = \theta_2 = \theta$ where θ is unknown, the likelihood function of θ is

$$L(\theta) = \prod_{i=1}^{n} (f(z_i;\theta_1))^{A_i} (S(z_i;\theta_1))^{B_i} \prod_{i=1}^{n} (f(M_i;\theta_2))^{C_i} (S(M_i;\theta_2))^{D_i} \tag{7}$$

$$L(\theta) = \prod_{i=1}^{n} ((1-\theta_1)^{z_i-1}\theta_1)^{A_i} ((1-\theta_1)^{z_i})^{B_i} \prod_{i=1}^{n} ((1-\theta_2)^{M_i-1}\theta_1)^{C_i} ((1-\theta_1)^{M_i})^{D_i} \tag{8}$$

Hence, the logarithm of the likelihood function is given by

$$\begin{aligned} \ln L(\theta) = &\sum_{i=1}^{n} \left[A_i\left((z_i-1)\ \ln(1-\theta) + \ln\theta\right) + B_i z_i \ln(1-\theta)\right] \\ &+ \sum_{i=1}^{n} \left[C_i\left((M_i-1)\ \ln(1-\theta) + \ln\theta\right) + D_i M_i \ln(1-\theta)\right] \end{aligned} \tag{9}$$

$$\begin{aligned} \frac{\partial(\ln L(\theta))}{\partial\theta} = &\sum_{i=1}^{n} \left[\left(\frac{A_i}{\theta} - \frac{A_i(z_i-1)}{1-\theta}\right) - \frac{B_i z_i}{1-\theta}\right] \\ &+ \sum_{i=1}^{n} \left[\left(\frac{C_i}{\theta} - \frac{C_i(M_i-1)}{1-\theta}\right) - \frac{D_i M_i}{1-\theta}\right] \end{aligned} \tag{10}$$

$$\sum_{i=1}^{n} A_i + \sum_{i=1}^{n} C_i = \theta\left[\left(\sum_{i=1}^{n} A_i z_i + \sum_{i=1}^{n} B_i z_i\right) + \left(\sum_{i=1}^{n} C_i M_i + \sum_{i=1}^{n} D_i M_i\right)\right]$$

The estimator of θ will be:

$$\begin{aligned} \hat{\theta} &= \frac{\sum_{i=1}^{n} A_i + \sum_{i=1}^{n} C_i}{\sum_{i=1}^{n}(A_i + B_i)z_i + \sum_{i=1}^{n}(C_i + D_i)M_i} \\ &= \frac{\frac{1}{2}\left(\sum_{i=1}^{n} \alpha_i^2\delta_i^2 + \sum_{i=1}^{n} \alpha_i\delta_i + \sum_{i=1}^{n} \beta_i^2\eta_i^2 + \sum_{i=1}^{n} \beta_i\eta_i\right)}{\sum_{i=1}^{n} \alpha_i^2\delta_i^2\, z_i + \sum_{i=1}^{n} \beta_i^2\eta_i^2\, M_i} \end{aligned} \tag{11}$$

2. Consistency and Asymptotic Normality of Estimators

In this section, the consistency and the asymptotic normality will be considered.

Theorem 1: $\hat{\theta}_1 \xrightarrow{a.s.} \theta_1$,

Proof: Since, $\{\alpha_i\delta_i, 1 \le i \le n\}$ are independently identical distributed variable, so

$$\frac{1}{n}\sum_{i=1}^{n}\alpha_i\delta_i \xrightarrow{a.s.} E(\alpha_i\delta_i)$$

$$E(\alpha_i\delta_i) = E(\alpha_i)E(\delta_i) = p[1-2(1-\theta_1)^{T_0}]$$

So

$$\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\alpha_i\delta_i \xrightarrow{a.s.} p[1-2(1-\theta_1)^{T_0}] \quad (12)$$

Similarly,

$$\frac{1}{n}\sum_{i=1}^{n}\alpha_i^2\delta_i^2 \xrightarrow{a.s.} E(\alpha_i^2\delta_i^2),$$

$$E(\alpha_i^2\delta_i^2) = p$$

So

$$\frac{1}{n}\sum_{i=1}^{n}\alpha_i^2\delta_i^2 \xrightarrow{a.s.} p, \quad (13)$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\alpha_i^2\delta_i^2 z_i \xrightarrow{a.s.} E(\alpha_i^2\delta_i^2 z_i)$$

$$E(\alpha_i^2\delta_i^2 z_i) = E(\alpha_i^2\delta_i^2)E(z_i) = p\left[\frac{1}{\theta_1}(1-(1-\theta_1)^{T_0}) - T_0(1-\theta_1)^{T_0}\right]$$

$$\frac{1}{n}\sum_{i=1}^{n}\alpha_i^2\delta_i^2 z_i \xrightarrow{a.s.} \frac{p}{\theta_1}(1-(1-\theta_1)^{T_0}) \quad (14)$$

Therefore, by using equations (12 and 14), and after simple calculations, it can be concluded that:

$$\hat{\theta}_1 \xrightarrow{a.s.} \theta_1 \quad (15)$$

By using the same method as we used to prove theorem1, we can prove that, $\hat{\theta}_2 \xrightarrow{a.s.} \theta_2$

Lemma: Let $T_{1n}, T_{2n}, \dots, T_{hn}$ be statistics such that as $n \to \infty$

$$\sqrt{n}(T_{1n} - \lambda_1, \dots \dots, T_{hn} - \lambda_h) \xrightarrow{d} N(0, \Sigma) \tag{16}$$

where $\Sigma = \left(\sigma_{ij}\right)_{h*h}, i = 1, \dots, h, j = 1, \dots, h.$

If $g(\lambda_1, \lambda_2, \dots, \lambda_h)$ is a function whose first derivatives all exist, then as $n \to \infty$

$$\sqrt{n}\big(g(T_{1n}, T_{2n}, \dots, T_{hn}) - g(\lambda_1, \lambda_2, \dots, \lambda_h)\big) \xrightarrow{d} N\left(0, \sum_{1}^{h}\sum_{1}^{h}\left(\frac{\partial g}{\partial \lambda_i}\frac{\partial g}{\partial \lambda_j}\right)\sigma_{ij}\right) \tag{17}$$

where $\partial g/\partial \lambda_i$ means $\partial g(\lambda_1, \lambda_2, \dots, \lambda_h)/\partial \lambda_i$ (Lawless, 2003)

Theorem 2: $\sqrt{n}\big(\hat{\theta}_1 - \theta_1\big) \xrightarrow{d} N(0, \sigma_1^2)$, where $\hat{\theta}_1$ defined as above.

Proof: assume that $Q_i = (\alpha_i\delta_i, \alpha_i^2\delta_i^2, \alpha_i^2\delta_i^2 z_i)^T$. let $\{Q_i, i \geq 1\}$ is iid variable.

Since,

$$E(\alpha_i\delta_i) = p[1 - 2(1-\theta_1)^{T_0}], E(\alpha_i^2\delta_i^2) = p, and\ E(\alpha_i^2\delta_i^2 z_i) = \frac{p}{\theta_1}[1 - (1-\theta_1)^{T_0}]$$

Then,

$$E(Q_i) = (p[1 - 2(1-\theta_1)^{T_0}], p, \frac{p}{\theta_1}[1 - (1-\theta_1)^{T_0}])^T \tag{18}$$

Let,

$$\Sigma = \mathrm{E}(Q_1 - E(Q_1))(Q_1 - E(Q_1))^T$$

By multivariate central limit theorem,

$$\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} Q_i - E(Q_1)) \xrightarrow{d} N(0, \Sigma)$$

where $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix}$

$$= \begin{pmatrix} var(\alpha_1\delta_1) & cov(\alpha_1\delta_1, \alpha_i^2\delta_i^2) & cov(\alpha_1\delta_1, \alpha_i^2\delta_i^2 z_i) \\ cov(\alpha_1\delta_1, \alpha_i^2\delta_i^2) & var(\alpha_i^2\delta_i^2) & cov(\alpha_i^2\delta_i^2, \alpha_i^2\delta_i^2 z_i) \\ cov(\alpha_1\delta_1, \alpha_i^2\delta_i^2 z_i) & cov(\alpha_i^2\delta_i^2, \alpha_i^2\delta_i^2 z_i) & var(\alpha_i^2\delta_i^2 z_i) \end{pmatrix}$$

$$\sigma_{11} = E(\alpha_1^2\delta_1^2) - \big(E(\alpha_1\delta_1)\big)^2 = p - p^2(1 - 2(1-\theta_1)^{T_0})^2$$

$$\sigma_{12} = \sigma_{21} = E(\alpha_1^3\delta_1^3) - (E\alpha_1\delta_1)(E\alpha_1^2\delta_1^2) = (p - p^2)(1 - 2(1-\theta_1)^{T_0})$$

$$\sigma_{13} = \sigma_{31} = E(\alpha_1^3\delta_1^3 z_1) - (E\alpha_1\delta_1)(E\alpha_1^2\delta_1^2 z_1)$$

$$= p(1-p)(1 - 2(1-\theta_1)^{T_0})(1 - (1-\theta_1)^{T_0})$$

$$\sigma_{22} = E(\alpha_1^4\delta_1^4) - \left(E(\alpha_1^2\delta_1^2)\right)^2 = p - p^2$$

$$\sigma_{23} = \sigma_{32} = E(\alpha_1^4\delta_1^4 z_i) - E(\alpha_1^2\delta_1^2)E(\alpha_1^2\delta_1^2 z_i) = \frac{p}{\theta_1}(1-p)(1 - (1-\theta_1)^{T_0})$$

$$\sigma_{33} = E(\alpha_1^4\delta_1^4 z_1^2) - (E\alpha_1^2\delta_1^2 z_1)^2$$

$$= p\left[\frac{(2-\theta_1)}{\theta_1^2}(1 - (1-\theta_1)^{T_0}) - T_0^2(1-\theta_1)^{T_0} - \frac{2T_0}{\theta}(1-\theta_1)^{T_0}\right] - \left(p(1 - (1-\theta_1)^{T_0})\right)^2$$

By using the above lemma, let

$$g(E_{1n}, E_{2n}, E_{3n}) = \frac{E_{1n} + E_{2n}}{2E_{3n}}, \tag{19}$$

where

$$E_{1n} = \frac{1}{n}\sum_{i=1}^{n}\alpha_i\delta_i\,, E_{2n} = \frac{1}{n}\sum_{i=1}^{n}\alpha_i^2\delta_i^2\,, E_{3n} = \frac{1}{n}\sum_{i=1}^{n}\alpha_i^2\delta_i^2 z_i,$$

$$\lambda_{1(\theta_1)} = E(E_{1n}) = p(1 - 2(1-\theta_1)^{T_0}), \qquad \lambda_{2(\theta_1)} = E(E_{2n}) = p,$$

$$\lambda_{3(\theta_1)} = E(E_{3n}) = \frac{p}{\theta_1}(1 - (1-\theta_1)^{T_0})$$

Assume that,

$$g(E_{1n}, E_{2n}, E_{3n}) = \hat{\theta}_1 \tag{20}$$

Then

$$g\left(\lambda_{1(\theta_1)}, \lambda_{2(\theta_1)}, \lambda_{3(\theta_1)}\right) = \frac{\lambda_{1(\theta_1)}, \lambda_{2(\theta_1)}}{2\lambda_{3(\theta_1)}} = \frac{p(1 - 2(1-\theta_1)^{T_0}) + p}{2\frac{p}{\theta_1}(1 - (1-\theta_1)^{T_0})} = \theta_1 \tag{21}$$

So,

$$\frac{\partial g\left(\lambda_{1(\theta_1)}, \lambda_{2(\theta_1)}, \lambda_{3(\theta_1)}\right)}{\partial \lambda_{1(\theta_1)}} = \frac{1}{2\lambda_{3(\theta_1)}} = \frac{\theta_1}{2p(1 - (1-\theta_1)^{T_0})} \tag{22}$$

$$\frac{\partial g\left(\lambda_{1(\theta_1)}, \lambda_{2(\theta_1)}, \lambda_{3(\theta_1)}\right)}{\partial \lambda_{2(\theta_1)}} = \frac{1}{2\lambda_{3(\theta_1)}} = \frac{\theta_1}{2p(1 - (1-\theta_1)^{T_0})} \tag{23}$$

$$\frac{\partial g\left(\lambda_{1(\theta_1)},\lambda_{2(\theta_1)},\lambda_{3(\theta_1)}\right)}{\partial \lambda_{3(\theta_1)}}=\frac{-\lambda_{1(\theta_1)}-\lambda_{2(\theta_1)}}{2\lambda_{3(\theta_1)}^2}=\frac{-\theta_1^2}{p(1-(1-\theta_1)^{T_0})} \tag{24}$$

Therefore,

$$\sqrt{n}\left(\hat{\theta}_1-\theta_1\right)=\sqrt{n}\left(g(E_{1n},E_{2n},E_{3n})-g\left(\lambda_{1(\theta_1)},\lambda_{2(\theta_1)},\lambda_{3(\theta_1)}\right)\right)\xrightarrow{d}N(0,\sigma_1^2) \tag{25}$$

where

$$\sigma_\theta^2=\frac{\partial^2 g}{\partial\lambda_{1(\theta_1)}^2}\sigma_{11}+2\frac{\partial^2 g}{\partial\lambda_{1(\theta_1)}\partial\lambda_{2(\theta_1)}}\sigma_{12}+2\frac{\partial^2 g}{\partial\lambda_{1(\theta_1)}\partial\lambda_{3(\theta_1)}}\sigma_{13}$$
$$=\frac{\partial^2 g}{\partial\lambda_{2(\theta_1)}^2}\sigma_{22}+2\frac{\partial^2 g}{\partial\lambda_{2(\theta_1)}\partial\lambda_{3(\theta_1)}}\sigma_{23}+\frac{\partial^2 g}{\partial\lambda_{3(\theta_1)}^2}\sigma_{33}$$

By using the same method as we used to provetheorem 2, we can prove

$$\sqrt{\boldsymbol{n}}\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{2}}-\boldsymbol{\theta}_{\boldsymbol{2}}\right)\xrightarrow{\boldsymbol{d}}\boldsymbol{N}\left(\mathbf{0},\boldsymbol{\sigma}_{\boldsymbol{2}}^{\mathbf{2}}\right)$$

3. Testing the Equality of Two Parameters and Confidence Interval for $\boldsymbol{\theta_1-\theta_2}$

In this section, hypothesis testing on the equality of two Geometric distributions under Type I censoring sample are constructed and its confidence interval are placed when some data are missing.

3.1 Hypothesis Testing for $\boldsymbol{\theta_1-\theta_2}$

The following hypotheses will be considered:

$$H_o:\theta_1-\theta_2=0\ ,\ H_1:\theta_1-\theta_2\neq 0.$$

First it is derived test statistics and discussed the limiting distribution of test statistics.

Assume that

$$\hat{\mathrm{p}}=\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}\delta_{\mathrm{i}}+\sum_{\mathrm{i}=1}^{\mathrm{n}}\eta_{\mathrm{i}}}{2\mathrm{n}}$$

where

$$E(\delta_i)=\frac{\sum_{i=1}^{n}\delta_i}{n}=p,E(\eta_i)=\frac{\sum_{i=1}^{n}\eta_i}{n}=p$$

By strong large number law,

$$\hat{p} \xrightarrow{a.s.} p, \hat{\sigma}_1^2 \xrightarrow{a.s.} \sigma_1^2, \hat{\sigma}_2^2 \xrightarrow{a.s.} \sigma_2^2$$

Therefore it is obtained the following result:

Theorem 3:

The test statistic is

$$\frac{\sqrt{n}\left[\hat{\theta}_1 - \hat{\theta}_2 - (\theta_1 - \theta_2)\right]}{\sqrt{\hat{\sigma}_{\theta_1}^2 + \hat{\sigma}_{\theta_2}^2}} \xrightarrow{d} N(0,1)$$

Under the-null hypothesis:

$$\frac{\sqrt{n}\left[\hat{\theta}_1 - \hat{\theta}_2\right]}{\sqrt{\hat{\sigma}_{\theta_1}^2 + \hat{\sigma}_{\theta_2}^2}} \xrightarrow{d} N(0,1) \tag{26}$$

Proof: By Slutsky's theorem, Theorem 1, Theorem 2, Theorem 3 can be proven.

3.2 Confidence Interval for $\theta_1 - \theta_2$

Let $\Delta\theta = \theta_1 - \theta_2$, in what follows it is discussed the confidence interval of Δ.

For $0 < \alpha < 1$, assume that ξ_α satisfies $\int_{-\infty}^{\xi_\alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \alpha$. For a given confidence level α , by Theorem 7, Theorem 8, so

$$P\left(\xi_{\frac{1-\alpha}{2}} \leq \frac{\sqrt{n}\left[\hat{\theta}_1 - \hat{\theta}_2 - \Delta\theta\right]}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}} \leq \xi_{\frac{1+\alpha}{2}}\right) \approx \alpha \tag{27}$$

Therefore it is obtained the α confidence interval of Δ:

$$\left(\hat{\theta}_1 - \hat{\theta}_2 - \xi_{\frac{1+\alpha}{2}} \frac{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}}{\sqrt{n}}, \hat{\theta}_1 - \hat{\theta}_2 + \xi_{\frac{1-\alpha}{2}} \frac{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}}{\sqrt{n}}\right) \tag{28}$$

References:

Lawless, J.F., (2003), "Statistical models and methods for lifetime data", John wiley &sons, 2nd Edition.

Zhao, Z., Wang, S., Wang, R., and Lil, I., (2009), "Parameter Estimation and Hypothesis Testing of Exponential Populations under Type Ⅰ Censoring Sample with Missing Data", Journal of Jilin University, vol. 47,No. 1, pp. 26 – 30.

A New Look at Bayesian Identification of Moving Average Models

Ayman A. Amin*

Abstract

In this paper we review the existing Bayesian identification techniques for the moving average (MA) models that can be classified as testing based identification and posterior mass function based identification. In order to improve the Bayesian identification of MA models, we present a new Bayesian identification method that is based on the posterior mass function of the MA models order. The main idea of our proposed Bayesian identification of MA models is that we first express the invertible MA model as an infinite autoregression in order to simplify the estimation of the unknown lagged errors using the ordinary least squares method (OLS), and based on the estimated errors we complete the Bayesian identification of MA models using the approximated posterior mass function of the model order. This method is simple and easy to apply, and simulation results show that its accuracy is comparable with other existing posterior mass function based identification and better than testing based identification.

Key Words: Time series model identification, Jeffreys' prior, Natural conjugate prior, Posterior mass function, Long autoregression.

1 Introduction

Time series models are widely used to fit and forecast time series data in many fields such as economics, finance and engineering. Time series analysis starts with the model identification and followed by the model estimation, model diagnostics check and finally model forecasting. Therefore, the model identification step is important since all other steps depend on its accuracy (Box et al., 2015). The identification of time series model means that the order of the time series model is unknown and needs to be specified (Amin, 2017c).

There are two Bayesian techniques proposed in the literature to identify the order of time series models. The first technique is testing based identification that assumes the model order is unknown constant with a known maximum, and the best order can be estimated based on a sequence of t-test of significance (Broemeling and Shaarawy, 1987). This Bayesian identification technique is applied to different time series models including autoregressive models (Broemeling and Shaarawy, 1988; Daif et al., 2003), autoregressive moving average models (Ali, 2003), and seasonal moving average models (El-Souda, 2008). The second technique is posterior mass function based identification that assumes the model order is a

* Assistant Professor of Statistics, Department of Statistics, Mathematics, and Insurance, Faculty of Commerce, Menoufia University, Egypt. Email: ayman.a.amin@gmail.com.

random variable with a known maximum, and its posterior mass function can be derived to select the order as a value with a maximum posterior probability. Following this idea, Diaz and Farah (1981) proposed a Bayesian method to identify the order of autoregressive models. Their work has been extended by researchers to different time series models, which include autoregressive moving average models (Fan and Yao, 2009), seasonal autoregressive models (Shaarawy and Ali, 2003), and multivariate autoregressive models (Shaarawy and Ali, 2008).

The Bayesian identification of the moving average (MA) models is complicated since the model errors are non-linear functions in the model coefficients. Accordingly, the errors sum of squares is non-quadric in the model coefficients and the likelihood function becomes analytically intractable, leading to non-standard posterior distribution. In order to address this problem, Broemeling and Shaarawy (1988) approximated the errors as linear functions in the coefficients by their non-linear least squares (NLS) estimates, and then they replaced the lagged errors of the model with their corresponding lagged residuals. This approach is adopted even for complicated time series models such double seasonal moving average models and double seasonal autoregressive moving average models (Amin, 2017a; Amin, 2017b; Amin, 2018b). Therefore, to accurately identify the MA models, Shaarawy et al. (2007) used fisrt the testing based identification technique to specify an initial value for the model order that is used to approximate the unknown lagged errors by their corresponding lagged residuals and then they returned to use the approximated posterior mass function based identification technique to improve the model identification.

It can be observed that this identification method proposed by Shaarawy et al. (2007) is iterative and computationally expensive, especially for higher order of MA models, because it depends on the nonlinear least squares method. Therefore, in this paper we address this problem by proposing a new method for Bayesian identification of MA models. The main idea of the proposed Bayesian identification of MA models is that we first exploit the fact that any invertible MA model can be expressed as an infinite autoregression in order to simplify the estimation of the unknown lagged errors using the ordinary least squares method (OLS), and based on the estimated errors we complete the Bayesian identification of MA models using the approximated posterior mass function of the model order. This method is simple and easy to apply, and results show that its accuracy is comparable with other existing methods.

The remainder of this paper is organized as follows. In Section 2 we present the background of the moving average time series models and related Bayesian concepts.In Section 3 summarize the existing Bayesian techniques for identification of MA models. In Section 4 we present our proposed Bayesian method for identifying MA models. In Section 5 we present the simulation study to evaluate the accuracy of our proposed Bayesian identification of MA models compared to other existing existing Bayesian identification techniques. Finally, we give the conclusions in Section 6.

2 Moving Average Models and Bayesian Concepts

Time series $\{y_t\}$ can be modeled by a moving average (MA) model of order q, simply denoted by MA(q), and written as (Box et al., 2015):

$$y_t = \theta_q(B)\varepsilon_t \quad (1)$$

where $\{\varepsilon_t\}$ is a sequence of independent and normally distributed errors with zero mean and variance σ^2, B is the backshift operator defined as $B^d x_t = x_{t-d}$, and $\theta_q(B)$ is the moving average polynomial with order q written as $\theta_q(B) = \left(1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q\right)$. The model (1) can be simplified and written as

$$y = X\beta + \varepsilon, \quad (2)$$

where $y = (y_1, y_2, \cdots, y_n)^T$, X is an $n \times q$ design matrix with the t^{th} row $X_t = \left(\varepsilon_{t-1}, \dots, \varepsilon_{t-q}\right)$, $\beta = \left(-\theta_1, \dots, -\theta_q\right)^T$ is the model coefficients, and $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)^T$.

It is worth noting that the design matrix X becomes a function of q when the MA model order is unknown. In this case we can assume that the model order q is a random variable with a known maximum value of k. The prior information about q can be represented in terms of a prior mass function $\zeta(q)$ that can have different forms such as uniform, i.e. $\zeta(q) = 1/k$, or geometric, i.e. $\zeta(q) = 0.5^q \forall q = 1,2,\dots,k$.

Bayesian analysis of time series models is based on Bayes' theorem that combines the prior distribution of the model parameters with the likelihood function of observed sample to get the posterior distribution.

Regarding the prior specification, we consider the natural conjugate and Jeffreys' priors. In case of the MA model with normally distributed errors, the natural conjugate prior is normal-gamma. Suppose $\beta \sim N_q(\mu_\beta, \sigma^2\Sigma_\beta)$ and $\sigma^2 \sim IG(\frac{\nu}{2}, \frac{\lambda}{2})$, the joint natural conjugate prior distribution of β and σ^2 is given by:

$$\zeta_n(\beta, \sigma^2) \propto (\sigma^2)^{-\left(\frac{\nu+q}{2}+1\right)} \exp\left\{-\frac{1}{2\sigma^2}\left[\lambda + \left(\beta - \mu_\beta\right)^T \Sigma_\beta^{-1}\left(\beta - \mu_\beta\right)\right]\right\}, \quad (3)$$

where $\mu_\beta, \Sigma_\beta, \nu$ and λ are hyperparameters need to be estimated.

Jeffreys' prior of β and σ^2 is given by:

$$\zeta_j(\beta, \sigma^2) \propto (\sigma^2)^{-1}, \sigma^2 > 0 \quad (4)$$

The likelihood function of the MA model (2) can be obtained by employing a straightforward random variable transformation from ε to y, and written as

$$L(\beta, \sigma^2, p|y) \propto (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\varepsilon^T\varepsilon\right\},$$
$$\propto (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right\}, \quad (5)$$

Multiplying the likelihood function (5) by each one of these two prior distributions results in the following joint posteriors. For the natural conjugate prior, the joint posterior of the model parameters β, σ^2 and q is:

$$\zeta_n(\beta, \sigma^2, q|y) \propto \zeta(q)(\sigma^2)^{-\left(\frac{n+\nu+q}{2}+1\right)} \exp\left\{-\frac{1}{2\sigma^2}\left[\lambda + \left(\beta - \mu_\beta\right)^T \Sigma_\beta^{-1}\left(\beta - \mu_\beta\right) + (y - X\beta)^T(y - X\beta)\right]\right\}. \quad (6)$$

For Jeffreys' prior, the joint posterior of β, σ^2 and q is:

$$\zeta_j(\beta, \sigma^2, q|y) \propto \zeta(q)(\sigma^2)^{-\left(\frac{n}{2}+1\right)} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right\}, \quad (7)$$

It is worth observing that the unknown lagged errors are part of the design matrices X which complicate the likelihood function and make it analytically intractable. As a result, computationally expensive numerical methods need to be introduced to obtain exact posterior of the model parameters as proposed by Monmohan (1983). Broemeling and Shaarawy (1987) approximated the likelihood function of the MA models by simply estimating the unknown lagged errors using the nonlinear least squares method and substituting these estimates in the likelihood function. In particular, they estimated the errors recursively by:

$$\hat{\varepsilon}_t = y_t + \sum_{i=1}^{q} \hat{\theta}_i \hat{\varepsilon}_{t-i}, \quad (8)$$

where $\hat{\theta}_1, \dots, \hat{\theta}_q$ are the nonlinear least squares estimates of the model coefficients obtained by minimizing

$$\varepsilon^T \varepsilon = SS(\theta_1, \dots, \theta_q), \quad (9)$$

with respect to $\theta_1, \dots, \theta_q$ over the invertibility regions. Accordingly, the main idea of this method is to search over the invertibility regions of the model coefficients and obtain their estimates as the values that minimize the errors sum of squares. Indeed, this makes the method computationally expensive to obtai the estimates of the unknown errors especially for higher orders of MA model, i.e. large values of q.

3 Existing Bayesian Identification of MA Models

There are two Bayesian techniques proposed in the literature to identify the order of time series models, which are testing based identification and posterior mass function based identification, and they are discussed in the following subsections.

3.1 Testing Based Identification of MA Models

The testing based identification technique for MA models is proposed by Broemeling and Shaarawy (1987) and can be summarized in two main steps. First, the model order is assumed to be unknown constant with a known maximum and accordingly the approximate posterior distribution of model coefficients is derived. Second, a sequence of t-test of significance is executed based on the marginal posterior of the model coefficients to specify the best value of the model order q.

Assuming the model order q is unknown constant with a known maximum k, from eqn (6) and using the NLS residuals (as given in eqn (8)) we can write the joint approximate posterior of β and σ^2 resulting from the natural conjugate prior as:

$$\zeta_n(\beta, \sigma^2|y) \propto (\sigma^2)^{-\left(\frac{n+\nu+k}{2}+1\right)} \exp\left\{-\frac{1}{2\sigma^2}\left[\lambda + \left(\beta - \mu_\beta\right)^T \Sigma_\beta^{-1}\left(\beta - \mu_\beta\right) + \left(y - \hat{X}\beta\right)^T\left(y - \hat{X}\beta\right)\right]\right\}. \quad (10)$$

where $\hat{X}$ is an $n \times k$ matrix with the t^{th} row:

$$\hat{X}_t = (\hat{\varepsilon}_{t-1}, \hat{\varepsilon}_{t-2}, \dots, \hat{\varepsilon}_{t-k}). \quad (11)$$

From this approximate joint posterior, it is easy to show that the approximate marginal posterior of the model coefficients β is a multivariate t distribution with degrees of freedom $v_n = (n + v)$ and location vector and dispersion matrix are respectively:

$$\mu_n = A_n^{-1}B_n and V_n = \frac{1}{v_n - 2}A_n^{-1}C_n, \quad (12)$$

where,

$A_n^{-1} = (\hat{X}^T\hat{X} + \Sigma_\beta^{-1})^{-1}$, $B_n = (\hat{X}^T y + \Sigma_\beta^{-1}\mu_\beta)$, and $C_n = \left[y^T y + \lambda + \mu_\beta^T\Sigma_\beta^{-1}\mu_\beta - B_n^T A_n^{-1}B_n\right]$.

Similarly for Jeffreys' prior, the joint posterior of β and σ^2 can be given as:

$$\zeta_j(\beta, \sigma^2|y) \propto (\sigma^2)^{-\left(\frac{n}{2}+1\right)}\exp\left\{-\frac{1}{2\sigma^2}\left(y - \hat{X}\beta\right)^T\left(y - \hat{X}\beta\right)\right\}. \quad (13)$$

We can easily show that the approximate marginal posterior of the model coefficients β is a multivariate t distribution with degrees of freedom $v_j = (n - k)$ and location vector and dispersion matrix are respectively:

$$\mu_j = (\hat{X}^T\hat{X})^{-1}\hat{X}^T y and V_j = \frac{1}{v_j - 2}(\hat{X}^T\hat{X})^{-1}C_j \quad (14)$$

where $C_j = \left[y^T y - y^T\hat{X}(\hat{X}^T\hat{X})^{-1}\hat{X}^T y\right]$.

The main property of the multivariate t distribution of a vector is that any single component of this vector has a univariate t distribution and the conditional distribution of any component given any other component has also a univariate t distribution. Using this marginal posterior of the parameters vector β and based on the property of the multivariate t distribution, we can do a backward elimination procedure to identify the best value of the model order q as follows:

1. Test $H_0: \theta_k = 0$ against $H_1: \theta_k \neq 0$ using the marginal posterior distribution of θ_k which is a univariate t distribution.
2. If H_0 is not rejected, test $H_0: \theta_{k-1} = 0$ against $H_1: \theta_{k-1} \neq 0$ using the conditional posterior distribution of θ_{k-1} given $\theta_k = 0$ which is also a univariate t distribution.
3. A sequence of t-test of significance is executed until the hypothesis $\theta_{q_0} = 0$ is rejected where $0 < q_0 \leq k$, which means the identified value for q is q_0.

3.2 Posterior Mass Function Based Identification of MA Models

The posterior mass function based identification assumes the model order q is a random variable with a known maximum k. Therefore, the marginal posterior mass function of the model order needs to be derived to select the order as a value with a maximum posterior probability (Amin, 2017d; Amin, 2018a). Since the lagged errors in the MA model are unknown, then the joint posteriors of β, σ^2 and q is analytically intractable. Therefore, an initial value can be assumed for q to use the NLS method to estimate the errors recursively (as given in eqn (8)) and then the joint posterior resulting from the natural conjugate prior can be approximated using the obtained residuals and given as:

$$\zeta_n(\beta,\sigma^2,q|y) \propto \zeta(q)(\sigma^2)^{-\left(\frac{n+\nu+q}{2}+1\right)}\exp\left\{-\frac{1}{2\sigma^2}\left[\lambda+\left(\beta-\mu_\beta\right)^T\Sigma_\beta^{-1}\left(\beta-\mu_\beta\right)+\left(y-\hat{X}\beta\right)^T\left(y-\hat{X}\beta\right)\right]\right\}. \quad (15)$$

By integrating out the parameters β and σ^2 in (15) and obtain the marginal posterior mass function of the model order q as:

$$\zeta_n(q|y) \propto \zeta(q)\left[\frac{|\Sigma_\beta^{-1}|}{|A_n|}\right]^{1/2}\left[y^Ty+\lambda+\mu_\beta^T\Sigma_\beta^{-1}\mu_\beta-B_n^TA_n^{-1}B_n\right]^{-\frac{n+\nu}{2}}\forall q=1,2,\ldots,k. \quad (16)$$

Where $A_n=(\hat{X}^T\hat{X}+\Sigma_\beta^{-1})$ and $B_n=(\hat{X}^Ty+\Sigma_\beta^{-1}\mu_\beta)$

For Jeffreys' prior, the joint posterior of β, σ^2 and q is given by :

$$\zeta_j(\beta,\sigma^2,q|y) \propto \zeta(q)(\sigma^2)^{-\left(\frac{n}{2}+1\right)}\exp\left\{-\frac{1}{2\sigma^2}\left(y-\hat{X}\phi\right)^T\left(y-\hat{X}\phi\right)\right\}, \quad (17)$$

Integrating out the parameters β and σ^2 in (17) results in the marginal posterior mass function of q as:

$$\zeta_j(q|y) \propto \zeta(q)\frac{\Gamma\left(\frac{n-q}{2}\right)}{\pi^{\frac{n-p}{2}}|\hat{X}^T\hat{X}|^{1/2}}\left[y^Ty-y^T\hat{X}(\hat{X}^T\hat{X})^{-1}\hat{X}^Ty\right]^{-\frac{n-q}{2}}\forall q=1,2,\ldots,k. \quad (18)$$

The main challenge in using the posterior mass function to identify the MA model is how to specify an initial value for the model order q to be able to use the NLS method to estimate the errors recursively and then approximate the joint posterior of the model parameters. Shaarawy et al. (2007) used the testing based identification technique to specify the initial value for q, however, this approach complicates the MA model identification and makes it computationally expensive.

4 Proposed Bayesian Identification of MA Models

Our proposed method for Bayesian identification of MA models is mainly based on the possibility of expressing the invertible MA model as a long finite autoregression to avoid the problem of specifying an initial value for q and also to avoid using the iterative NLS method to estimate the errors and approximate the joint posterior. To explain our idea, consider the MA(1) model:

$$y_t=-\theta_1\varepsilon_{t-1}+\varepsilon_t. \quad (19)$$

This model can be expressed as an infinite autoregression (Box et al. 2015) as:

$$y_t=\pi_1y_{t-1}+\pi_2y_{t-2}+\pi_3y_{t-3}+\cdots+\varepsilon_t. \quad (20)$$

However, one can approximate this inifinite autoregression by a long finite autoregression of order L, assuming all the values of $\pi_i s$ are zero for lags beyond L (Koreish and Pukkila, 1990), and it can be written as:

$$y_t=\pi_1y_{t-1}+\pi_2y_{t-2}+\cdots+\pi_Ly_{t-L}+\varepsilon_t. \quad (21)$$

From (19) and (21), with some manipulations we can find the values of $\pi_i s$ in terms of the model coefficient θ_1 as: $\pi_i=\theta_1^i,\forall i=1,\ldots,L$. By substituting these values of $\pi_i s$ in eqn

(21), we get:

$$\begin{aligned} y_t &= -\theta_1 y_{t-1} + \sum_{i=2}^{L} (-\theta_1) y_i + \varepsilon_t, \\ &= -\theta_1 \left[y_{t-1} - \sum_{i=2}^{L} (-\theta_1)^{i-1} y_{t-i} \right] + \varepsilon_t, \\ &= -\theta_1 [y_{t-1} - \sum_{i=2}^{L} \pi_{i-1} y_{t-i}] + \varepsilon_t \end{aligned} \quad (22)$$

The quantity $[y_{t-1} - \sum_{i=2}^{L} \pi_{i-1} y_{t-i}]$ is just ε_{t-1}, which proves that the MA(1) model can be expressed as a long finite autoregression of order L, and in the same way we can prove that for higher order of MA models.

Accordingly, our proposed method for Bayesian identification of MA models can be summarized as follows.

1. First, we express the MA model as a long autoregressive model (LAR) of order L as:

$$y_t = \pi_1 y_{t-1} + \pi_2 y_{t-2} + \cdots + \pi_L y_{t-L} + \varepsilon_t, \quad (23)$$

and then we use the OLS method to estimate the LAR(L) model coefficients, $\hat{\pi}_i$'s, and obtain the residuals, $\hat{\varepsilon}_t$'s, as consistent estimates for the unknown lagged errors.

2. Second, we replace the unknown lagged errors with the obtained OLS residuals in the likelihood function (5) to approximate it and the resulting approximate posterior density of the model parameters β, σ^2 and q will be analytically tractable and it is given as in eqn (15) for the natural conjugate prior and as in eqn (17) for Jeffreys' prior.

3. Third, we integrate out the parameters β and σ^2 to obtain the approximate marginal posterior mass function of q and it is given as in eqn (16) for the natural conjugate prior and as in eqn (18) for Jeffreys' prior, and then we select the order as a value with a maximum posterior probability.

The main challenge in our proposed method is the specification of the order L of LAR model. Initially, we have tried to use several information criteria, such as Akaike's information criterion (AIC), to specify the order L but unfortunately we got unacceptable results. For the sake of illustration, we consider simple simulated examples, MA(1) model with $\theta = 0.5$ and 0.8. For all the examples we consider $n = 100$. Using the AIC to determine the order L and other arbitrarily specified orders, i.e. $L = 5,10,15,20$, we apply our proposed method, assuming Jeffreys' prior for the model parameters, and our Bayesian estimates are presented in Table (1).

	$\theta = 0.5$					$\theta = 0.8$				
	AIC	Arbitrarily Specified				AIC	Arbitrarily Specified			
L	4	5	10	15	20	6	5	10	15	20
$E(\theta\|y)$	0.515	0.504	0.504	0.537	0.519	0.780	0.756	0.782	0.795	0.818
$V(\theta\|y)$	0.011	0.011	0.012	0.013	0.015	0.012	0.012	0.012	0.013	0.016

Table 1: Simulated examples for LAR model specification.

From this table, we can observe that the Bayesian estimates resulting from using the order L specified by the AIC is not better than those arbitrarily specified. These results confirm that the AIC and other information criteria are not a suitable method to specify the order of

LAR model. In general, in our proposed method we need the order L to be large enough to enable the LAR model approximates adequately the unknown lagged errors, and in the same time it has to be less than the number of observations of the given time series, i.e $L < n$. Extending the simulation with considering L as a function of $\sqrt{n}$, we get the results in Table (2) that reveal using $L = \sqrt{n}$ can be a good choice which is consistent with the results of Koreish and Pukkila (1990) in the non-Bayesian domain.

n	$\theta = 0.5$	L					
		$0.5\sqrt{n}$	$1.0\sqrt{n}$	$1.5\sqrt{n}$	$2.0\sqrt{n}$	$2.5\sqrt{n}$	$3.0\sqrt{n}$
50	Mean	0.495	0.501	0.503	0.498	0.491	0.486
	Std. Dev.	0.151	0.157	0.165	0.172	0.181	0.191
100	Mean	0.499	0.502	0.497	0.499	0.499	0.499
	Std. Dev.	0.104	0.107	0.111	0.114	0.118	0.122
150	Mean	0.496	0.494	0.495	0.496	0.498	0.497
	Std. Dev.	0.084	0.087	0.089	0.091	0.094	0.096
200	Mean	0.495	0.495	0.496	0.496	0.497	0.496
	Std. Dev.	0.073	0.074	0.076	0.078	0.079	0.081
250	Mean	0.491	0.491	0.490	0.493	0.493	0.495
	Std. Dev.	0.065	0.066	0.068	0.069	0.070	0.071
300	Mean	0.494	0.493	0.493	0.493	0.493	0.494
	Std. Dev.	0.059	0.060	0.061	0.063	0.064	0.065

Table 2: Simulated examples for LAR model specification with L is a function of $\sqrt{n}$.

5 Simulation Study

In this section we present a simulation study to evaluate the accuracy of the proposed Bayesian identification method for MA models, compared to other existing Bayesian identification methods, for several simulated time series data with different sample size, different model orders, and different values of the model coefficients. To run the simulations, we generate 1,000 time series of size n from 50 to 400 from MA models with order one, where θ = 0.3, 0.5, and 0.8, and with order two, where (θ_1, θ_2) = (0.2, 0.6), (0.5, 0.4) and (0.9, 0.4).

Once the time series datasets are generated from these MA models, the Bayesian analysis is performed by assuming Jeffreys' prior for the model parameters β and σ^2, and employing three prior distributions for the model order q given as:

$$\zeta_1(q) = \frac{1}{k}, \forall q = 1,2,\dots,k(\text{Uniformprior})$$
$$\zeta_2(q) = 0.5^q, \forall q = 1,2,\dots,k(\text{Geometricprior})$$
$$\zeta_3(q) = \frac{k-q+1}{k+1}, \forall q = 1,2,\dots,k(\text{Arithmeticprior}) \quad (24)$$

In order to apply the testing based identification technique for each generated time series, we execute a sequence of significance based on the marginal posterior of the model coefficients with assuming the maximum value of the model order is $k = 4$ and specify the

best value of the model order q. On the other hand, to apply the posterior mass function based identification we compute the posterior mass functions of the MA model order, $\zeta_1(q|y)$, $\zeta_2(q|y)$, and $\zeta_3(q|y)$ resulting from the employed priors in eqn (24) respectively with assuming the maximum value of the model order is $k = 4$, and then we identify the model order as a value with a maximum posterior probability. For all simulated time series, we compute the percentage of correctly identified models by comparing the identified order with the true value of q used to generate the time series. Results of the simulation study are presented in Tables (3) and (4).

n	Method						
	Testing [1]	NLS-based PMF [2]			LAR-based PMF [3]		
		$\zeta_1(p\|y)$	$\zeta_2(p\|y)$	$\zeta_3(p\|y)$	$\zeta_1(p\|y)$	$\zeta_2(p\|y)$	$\zeta_3(p\|y)$
$\theta = 0.3$							
50	80.7	73.6	86.8	82.3	73.6	90.0	83.9
100	83.0	84.5	92.2	89.9	83.1	93.6	88.9
200	82.1	88.2	93.4	91.5	88.6	94.5	92.1
300	84.4	91.8	95.7	93.2	91.6	96.0	93.7
400	85.2	93.2	96.2	95.2	93.8	96.4	95.4
$\theta = 0.5$							
50	75.9	71.5	83.6	78.4	68.3	87.7	81.1
100	78.7	82.2	90.4	87.7	80.9	90.4	86.9
200	78.3	86.1	91.6	88.9	85.8	92.7	89.8
300	80.2	88.9	93.3	91.1	88.7	94.2	91.7
400	81.1	91.0	95.0	93.5	91.5	95.2	93.6
$\theta = 0.5$							
50	70.7	68.0	77.4	74.4	59.2	82.0	73.3
100	69.3	73.9	83.1	80.0	71.9	86.3	81.8
200	70.7	79.1	86.2	81.8	77.2	88.0	83.7
300	73.5	83.2	89.6	85.9	83.5	90.2	86.6
400	70.7	83.4	91.0	87.4	85.1	91.6	89.2

[1] Testing based identification.

[2] Posterior mass function based identification using NLS to estimate errors.

[3] Posterior mass function based identification using LAR to estimate errors.

Table 3: Percentage of correctly identified models for MA(1).

From the simulation results we can observe some remarks. Firs, the larger sample size the higher percentage of correctly identified models is obtained. Second, the identification results obtained from the posterior mass function techniques are better than those obtained from the testing based technique especially for large sample size. Third , the employed priors for the model order result in different posteriors, and their impact can be observed in the percentage of correctly identified models since the percentage of correctly identified models for the geometric and arithmetic priors in most of the cases is higher than that obtained for the uniform prior. Fourth, our proposed Bayesian identification method, LAR-based PMF, in all of the cases is better than the testing based technique and at least is comparable with the other

posterior mass function technique, NLS-based PMF, proposed by Shaarawy et al. (2007).

n	Method						
	Testing	NLS-based PMF			LAR-based PMF		
		$\zeta_1(p\|y)$	$\zeta_2(p\|y)$	$\zeta_3(p\|y)$	$\zeta_1(p\|y)$	$\zeta_2(p\|y)$	$\zeta_3(p\|y)$
$\theta_1 = 0.2$ and $\theta_2 = 0.6$							
50	83.1	79.3	85.8	84.4	70.7	86.0	83.0
100	84.8	84.2	90.6	89.7	82.1	92.4	89.2
200	83.7	88.1	91.9	90.7	86.0	93.3	91.5
300	84.9	89.7	93.9	92.6	88.7	95.5	93.8
400	83.3	89.6	94.8	93.3	90.7	95.9	94.1
$\theta_1 = 0.5$ and $\theta_2 = 0.4$							
50	58.6	61.5	61.5	63.3	61.5	62.7	67.2
100	81.8	82.0	86.3	85.3	79.6	84.3	84.7
200	85.0	88.8	92.2	91.1	85.3	92.3	91.2
300	86.7	90.7	94.7	93.7	89.4	94.8	93.3
400	84.5	90.6	95.1	93.6	90.1	95.5	93.8
$\theta_1 = 0.9$ and $\theta_2 = -0.3$							
50	26.1	28.3	28.0	29.3	36.1	30.3	36.4
100	44.5	44.5	48.3	48.2	52.1	47.0	51.6
200	67.5	71.7	72.6	73.7	74.3	71.6	74.4
300	77.8	83.8	84.1	84.6	84.7	85.3	85.8
400	80.4	85.9	90.0	89.0	87.4	92.4	91.2

Table 4: Percentage of correctly identified models for MA(2).

6 Conclusion

In this paper we first reviewed and summarized the existing Bayesian identification techniques for the moving average (MA) models that can be classified as testing based identification and posterior mass function based identification. The posterior mass function based identification proposed for the MA models is iterative and computationally expensive, therefore to improve the Bayesian identification of these models, we presented a new Bayesian identification method that is based on the posterior mass function of the model order. The main idea of the proposed identification method is that we express the invertible MA model as an infinite autoregression in order to simplify the estimation of the unknown lagged errors using the ordinary least squares method (OLS), and based on the estimated errors we complete the Bayesian identification of MA models using the approximated posterior mass function. We used a large number of Monte Carlo simulations to evaluate the accuracy of the proposed method compared to other existing techniques, and the simulation results show that the accuracy of our proposed Bayesian identification method is comparable with other existing posterior mass function based identification and better than testing based identification. Future work may be an extension to multivariate moving average models.

References

Ali, S. S. (2003). Bayesian Identification of ARMA Models. Unpublished Ph.D. Dissertation, Department of statistics, Faculty of Economics and Political Science, Cairo University, Egypt.

Amin, A. (2017a). Bayesian inference for double seasonal moving average models: A Gibbs sampling approach. Pakistan Journal of Statistics and Operation Research, 13(3), 483–499.

Amin, A. (2017b). Gibbs Sampling for Double Seasonal ARMA Models. In Proceedings of the 29th Annual International Conference on Statistics and Computer Modeling in Human and Social Sciences. Faculty of Economics and Political Science, Cairo University, Egypt.

Amin, A. (2017c). Sensitivity to Prior Specification in Bayesian Identification of Autoregressive Time Series Models. *Pakistan Journal of Statistics and Operation Research,* 13(4), 699-713.

Amin, A. (2017d). Identification of Double Seasonal Autoregressive Models: A Bayesian Approach. In Proceedings of the 52nd Annual International Conference of Statistics, Computer Science and Operations ResearchAt: ISSR, Cairo University, Egypt.

Amin, A. (2018a). Identification of Double Seasonal Autoregressive Models: A Bayesian Approach. In Proceedings of the 52nd Annual International Conference of Statistics, Computer Science and Operations Research. ISSR, Cairo University, Egypt.

Amin, A. (2018b). Bayesian Inference for Double SARMA Models. *Communications in Statistics: Theory and Methods*, 47 (21), 5333-5345.

Box, G., Jenkins, G., Reinsel, G., and Ljung, G. (2015). Time series Analysis, Forecasting and control. Fifth Edition, John Wiley & Sons.

Broemeling, L. and Shaarawy, S. (1987). Bayesian Identification of Time Series. The 22nd Annual Conference in Statistics , Computer Science and Operation Research, Institute of Statistical Studies and Research, Cairo, Egypt, Vol.1,pp.146-159.

Broemeling, L. and Shaarawy, S. (1988). Time Series: A Bayesian Analysis in Time Domain. Studies in Bayesian Analysis of Time Series and Dynamic Models, edited by J. Spall, Marcel Dekker Inc, New York, pp 1-21.

Dias, J. and Farah, J. L. (1981). Bayesian Identification of Autoregressive Process. Presented at the 22nd NBER-NSC Seminar on Bayesian Inference in Econometrics.

Daif, A. , Soliman, E. and Ali, S.(2003):On Direct and Indirect Bayesian Identification of Autoregressive Models, The 15 th Annual Conference in Statistics and Computer Modeling in Human and Social Sciences, Faculty of Economics and Political Science , Cairo University.

El-Souda, R.M. (2008). Bayesian Identification for Seasonal Time Series models. Unpublished PH.D Dissertation, Department of Statistics, Faculty of Economics and Political Science, Cairo University, Egypt.

Fan, C. and Yao, S. (2009). Bayesian approach for arma process and its application. International Business Research, 1(4):49–55.

Koreish, S. and Pukkila, T. (1990). Linear Methods for Estimating ARMA and Regression Models with Serial Correlation. Communications in Statistics - Simulation and Computation.

Monahan, J.F.(1983): Fully Bayesian analysis of ARMA time series models , J. of Econometrics, Vol.21, pp.307-331.

Shaarawy, S. M., Soliman, E. A. and Ali, S. S. (2007): Bayesian Identification of Moving Average Models, Communications in Statistics - Theory and Methods, 36 (12): 2301-2312.

Shaarawy, S. and Ali, S.(2003). Bayesian Identification of Seasonal Autoregression Models. Communications in Statistics – Theory and Methods, Vol.32, Issue 5, pp 1067-1084.

Shaarawy, S. and Ali, S. (2008). Bayesian Identification of Multivariate Autoregressive Models. Communications in Statistics –Theory and Methods, Vol. 37, Issue 5 , pp 791-802.

Ridge Estimators for the Negative Binomial Regression Model with Application

El-Housainy A. Rady[1], Mohamed R. Abonazel[2], & Ibrahim M. Taha[3*]

Abstract

The common method for modeling count data in case of over-dispersed data is the negative binomial (NB) regression model. The NB regression model is estimated using maximum likelihood (ML) method. The ML method is very sensitive to high inter-correlation among the explanatory variables; which is commonly referred as multicollinearity problem. Therefore, we present some ridge estimators for the NB ridge regression to remedy the problem of instability of the traditional ML method and increase the efficiency of estimation. Finally, a real dataset application is conducted to investigate the performance of the ridge estimators and the traditional ML method. The results show that the ridge regression method outperforms ML estimator for all different ridge parameters considered in this study, since the ridge regression estimates have smaller standard errors than ML estimator in the application.

Key words: Multicollinearity – Ridge Regression – GLMs – Negative Binomial Regression – IRLS

1. Introduction

Multicollinearity is a high inter-correlation among two or more explanatory variables in the regression models which can seriously distort the ML estimates. It has been well known that high linear dependency among covariates in any regression model including not only traditional regression models but generalized linear models can cause frustrating statistical problems; such as higher inflated variances, lowered power in prediction and even incorrect signs to the estimation's results.

[1,2] Department of Applied Statistics and Econometrics, Institute of Statistical Studies and Research, Cairo University, Egypt.
[3] Department of Mathematics, Statistics, and Insurance, Sadat Academy for Management Sciences, Tanta Branch, Egypt. *E-mail : ibrahimaboalazm@gmail.com

We can also define multicollinearity through the concept of orthogonality; when the predictors are orthogonal or uncorrelated, all eigenvalues of the design matrix are equal to one and the design matrix is full rank. If at least one eigenvalue is different from one, especially when equal to zero or near zero; then non-orthogonality exists, meaning that multicollinearity is present.

There are some sources of multicollinearity that impact the analysis, the corrections, and the interpretation of the linear model; some of these sources are: data collection, physical constraints, over-defined model, model choice or specification, and outliers. See Montgomery et al. (2012) for details. Multicollinearity has several effects, such as:

- High variance of estimates may reduce the efficiency of estimation.
- Multicollinearity can result in coefficients appearing to have the wrong sign.
- Estimates of coefficients may be sensitive to particular sets of sample data.
- Some variables may be dropped from the model although, they are important in the population.
- The coefficients are sensitive to the presence of small number inaccurate data values. For more details see Gujarati (2009).

Several procedures for detecting multicollinearity problem in linear regression have been proposed in the literature. Some of these methods are:

(i) **Examination of the Correlation Matrix:** A simple and an efficient method for detecting multicollinearity is to calculate the correlation coefficients between any two of the explanatory variables. A high value of the correlation between two variables may indicate that the variables are collinear. This method is easy, but it cannot produce a clear estimate of the degree of multicollinearity if the correlation coefficients are greater than 0.80 or 0.90 then this is an indicator of high multicollinearity.

(ii) **Variance Inflation Factor (VIF):** The VIF quantifies the severity of multicollinearity in the regression model. Let R_j^2 denotes the coefficient of determination, when X_j is regressed on all other predictor variables in the model. The VIF is given by:

$$\text{VIF} = \frac{1}{1-R_j^2} \ , \qquad j = 1,2,\dots,p,$$

where p is number of variables. The VIF provides an index that measures how much the variance of an estimated regression coefficient is increased because of the multicollinearity. As per practical experience, if any of the VIF values exceeds 5 or 10, it is an indication that the associated regression coefficients are poorly estimated because of multicollinearity.

As a remedy to this problem, caused by multicollinearity some methods have been proposed to combat this problem, some of these methods are;

i.	Collecting Additional Data	ii.	Model Respecification
iii.	Drop one of the correlated variables	iv.	Partial Least Squares (PLS)
v.	Principal Component (PC) Regression	vi.	Biased Estimation

This paper is organized as follows: section 2 presents a background about generalized linear models; especially the negative binomial model. In section 3 the ridge regression method is discussed, a real dataset in section 4 is used to evaluate the performance of the new method. In Section 5 we present the final conclusion.

2. Background

In this section we give a background about generalized linear models, describe the negative binomial model, outline the assumptions of the model, and how the parameters are estimated.

2.1 Generalized Linear Models

Generalized linear models (GLMs) by McCullagh and Nelder (1989) represent a class of regression models appropriate to investigate the effect of input variables over non-normal response variables. The GLM model is based on probability distributions with unknown location parameter (Θ) that belongs to the exponential family. The most important distributions in this family are normal, binomial, Poisson, gamma, negative binomial, and exponential. The exponential family probability density function is usually described as:

$$\mathrm{f}\,(y_i; \Theta_i, \phi) = \exp\left\{\frac{(y_i\Theta_i - b(\Theta_i))}{\alpha_i(\phi)} + c(y_i, \phi)\right\}, \quad i = 1,2, \dots, n \qquad (1)$$

where

y is the response variable

Θ_i is the canonical parameter or the natural parameter

$b(\Theta_i)$ is the cumulant from which the mean and variance functions are derived; cumulants are set of quantities that provide an alternative to the moments of the distribution.

$\alpha_i(\phi)$ is the scale parameter, set to one in discrete and count models

$c(y_i, \phi)$ is the normalization term, guaranteeing that the probability function sums to unity.

GLMs are structured by three components: (i) Random component; which defines the probability distribution of the response variable y, (ii) Systematic component η; which is the linear predictor that defines the structure of the input variables, and (iii) Link function; which describes the functional relationship between the systematic component and the expected value for the random component (i.e., the mean of response variable y), and the variance of y is a function of μ.

The systematic component comprising the regression model (η) is the linear combination of input variables, and it may be written by a $\mathrm{g}(\cdot)$ function, called link

function, it has to be a monotone, differentiable function; it describes the functional relationship between the μ mean and the linear predictor (η)

$$g(\mu) = \eta = x'\beta,$$

where β is a vector of the unknown coefficients and X is a matrix of the input variables.

2.2 Negative Binomial Regression Model

The negative binomial (NB) regression model is one basic framework for count data analysis. This model has found a widespread use in the fields of health, social, economic and physical sciences when the dependent variable y_i comes in the form of non-negative integers or counts. It has three basic assumptions; on the conditional distribution of the dependent variable, on the specification of the mean parameter, and on the independence of the distribution for all observations. There are twenty-two different versions of NB model were mentioned by Hilbe (2011), however we rely on the traditional NB model by Cameron and Trivedi (1986), with mean μ and variance function $\mu + \theta\mu^2$.

The traditional negative binomial model has the same distributional assumptions as the Poisson distribution, with the exception that it has a second parameter "the dispersion parameter" which provides for a wider shape to the distribution of counts than is allowed under Poisson assumptions. The Poisson assumption of equi-dispersion means that the values of the mean and variance are the same. For the negative binomial, two parameters affect the variance over that of the mean "the dispersion parameter (θ)" and square of the mean (μ^2), with greater values of the negative binomial mean come much greater values of the variance. So, in case of over-dispersion the negative binomial regression model is favorable than the Poisson regression model.

2.2.1 Assumptions of Negative Binomial Model

The general assumptions involved in negative binomial regression model are as follows;

1. The response y, is a count consisting of nonnegative integers.
2. As the value of μ increases, the probability of 0 counts decreases.

3. y must allow for the possibility of 0 counts.
4. The fitted or predicted variable μ, is the expected mean of the distribution of y.
5. A foremost goal of NB regression is to model data in which the value of the variance exceeds the mean, or the observed variance exceeds the expected variance.
6. A well-fitted NB model has a dispersion statistic approximating 1.0 and an AIC/BIC and log-likelihood statistic less than alternative count models.
7. The model is not misspecified.
8. The number of predicted counts is approximately the same as the number of observed counts across the distribution of y.

For negative binomial models we can describe the negative binomial probability mass function (PMF) as the probability of observing y failures before the r^{th} success in a series of Bernoulli trials. Under such a description r is a positive integer. Also, π is rather thought of as the probability of r successes and y the number of failures before the r^{th} success.

It should be emphasized that a negative binomial regression model has a negligible tie to how the underlying PMF is derived. When such a model is being used to accommodate Poisson over-dispersion, or to estimate predicted counts, it matters little how many failures have occurred before a specific number of successes. The probability mass function of the negative binomial distribution maybe written as

$$\mathrm{f}(y;\pi,r) = \binom{y_i + r - 1}{r - 1} \pi_i^r (1-\pi_i)^{y_i}.$$

Converting the negative binomial PMF into exponential family form results in:

$$\mathrm{f}(y;\pi,r) = \exp\left\{y_i \log(1-\pi_i) + r\log(\pi_i) + \log\binom{y_i + r - 1}{r - 1}\right\}.$$

The mean and variance of NB after re-parameterization are respectively as follows:

$$b'(\Theta_i) = \frac{r\,(1-\pi_i)}{\pi_i} = \mu_i\,,$$

$$b''(\Theta_i) = \frac{r\,(1-\pi_i)}{\pi_i^2} = \mu_i(1+\theta\mu_i),$$

where $\theta = 1/r$. Thus, the NB model adds a quadratic term $\theta\mu^2$ to the variance of Poisson to account for extra-Poisson variation or over-dispersion. For this reason, θ is known as the dispersion parameter. Given the defined values of μ and θ, we may re-parameterize the negative binomial PMF such that

$$\mathrm{f}(y;\mu,\theta) = \begin{pmatrix} y_i + \left(\frac{1}{\theta}\right) - 1 \\ \left(\frac{1}{\theta}\right) - 1 \end{pmatrix} \left(\frac{1}{1+\theta\mu_i}\right)^{\frac{1}{\theta}} \left(\frac{\theta\mu_i}{1+\theta\mu_i}\right)^{y_i}.$$

The likelihood function can be derived as a Poisson–gamma mixture as;

$$L(\mu;y,\theta) = \prod_{i=1}^{n} \exp\left\{ y_i \log\left(\frac{\theta\mu_i}{1+\theta\mu_i}\right) - \frac{1}{\theta}\log(1+\theta\mu_i) + \log\Gamma\left(y_i + \frac{1}{\theta}\right) - \log\Gamma(y_i+1) - \log\Gamma\left(\frac{1}{\theta}\right)\right\}.$$

The log-likelihood is obtained by taking the natural log of both sides of the equation. As with the Poisson model, the function becomes additive rather than multiplicative.

$$\mathcal{L}(\mu;y,\theta) = \sum_{i=1}^{n} y_i \log\left(\frac{\theta\mu_i}{1+\theta\mu_i}\right) - \frac{1}{\theta}\log(1+\theta\mu_i) + \log\Gamma\left(y_i + \frac{1}{\theta}\right) - \log\Gamma(y_i+1) - \log\Gamma\left(\frac{1}{\theta}\right).$$

In most cases the conditional variance exceeds the conditional mean, which is commonly referred to as over dispersion. This often comes from neglected or unobserved heterogeneity that is inadequately captured by the explanatory variables in the conditional mean function. So, we usually allow the Poisson conditional mean to be randomly distributed as gamma.

Hence, the dependent variable is generated using a stochastic mechanism that corresponds to a Poisson–gamma mixture instead of a Poisson distribution, which leads to a marginal NB distribution. Then the Poisson regression model should be replaced by the NB regression model, since the standard errors of the slope coefficients otherwise will be underestimated.

$$y_i = \mathrm{E}(y_i|\, x_i) + \varepsilon_i,$$

This model is attractive because it manages to handle data that is over-dispersed since it allows for random variation in the Poisson conditional mean H_i by letting

$$\mathrm{E}(y_i|\, x_i) = H_i = z_i\mu_i, \qquad (2)$$

where $\mu_i = \exp(x_i'\beta)$, x_i is the i^{th} row of X which is a $n \times (p+1)$ data matrix with p explanatory variables, β is a $(p+1) \times 1$ vector of coefficients and z_i a random variable that is $\Gamma\left(\frac{1}{\delta}, \frac{1}{\delta}\right)$ distributed. The model allows μ_i to depend on covariates through the relationship

$$\mathrm{h}(\mu_i) = \log\left(\frac{\theta\mu_i}{1+\theta\mu_i}\right) = x_i'\beta.$$

2.2.2 Maximum Likelihood Estimation

This model is usually estimated by the maximum likelihood (ML) estimator which is found by maximizing the log-likelihood function

$$\mathcal{L}(\theta, \beta; y) = \sum_{i=1}^{n}\left\{\left(\sum_{j=0}^{y_i-1}\log\left(j+\frac{1}{\theta}\right)\right) - \log y_i! - \left(y_i+\frac{1}{\theta}\right)\log(1+\theta\mu_i) + y_i\log(\theta) + y_i\log\mu_i\right\}, \qquad j = 1,2,\dots,p.$$

Then the ML can be obtained through the following score function:

$$S(\beta) = \frac{\partial\mathcal{L}(\theta,\beta)}{\partial\beta} = \sum_{i=1}^{n}\frac{y_i-\mu_i}{1+\theta\mu_i}x_i. \qquad (3)$$

Equating Equation (3) to zero and solving, we can see that it is nonlinear in β, the solution is found by applying the method of Fisher scoring;

$$\beta^r = \left(\beta^{(r-1)}\right) + \mathrm{I}^{-1}\left(\beta^{(r-1)}\right) S\left(\beta^{(r-1)}\right),$$

where I is the information matrix, then $S\left(\beta^{(r-1)}\right)$ and $\mathrm{I}^{-1}\left(\beta^{(r-1)}\right)$ are $S(\beta)$ and $\mathrm{I}^{-1}(\beta)$ evaluated at $\beta^{(r-1)}$ respectively;

$$\mathrm{I}^{-1}(\beta) = \left[-E\left(\frac{\partial^2\mathcal{L}(\theta,\beta)}{\partial\beta\partial\beta'}\right)\right]^{-1} = \left[E\left(\sum_{i=1}^{n}x_i x_i'\theta\frac{\mu_i(\theta+y_i)}{(\theta+\mu_i)^2}\right)\right]^{-1}.$$

In the final step, the value of β that maximizes the likelihood function is obtained as

$$\hat{\beta}_{\mathrm{ML.NB}} = \left(X'\widehat{W}_{\mathrm{NB}}X\right)^{-1}(X'\widehat{W}_{\mathrm{NB}}\hat{z}), \qquad (4)$$

where $\hat{\beta}_{\mathrm{ML.NB}}$ is the vector of estimated parameters of negative binomial model using maximum likelihood method, and $\widehat{W}_{\mathrm{NB}}$ is a weighting matrix where the off-diagonal elements are zeros and the i^{th}diagonal element is equal to $\hat{\mu}_i$

$$\widehat{W}_{\mathrm{NB}} = \mathrm{diag}\left(\frac{\hat{\mu}_i}{(1+\hat{\mu}_i)/\hat{\theta}}\right),$$

and $\hat{z}$ is a vector where the i^{th} element equals

$$\hat{z} = \log(\hat{\mu}_i) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}.$$

The ML estimator of β is normally distributed with asymptotic mean vector $E\left(\hat{\beta}_{\mathrm{ML}}\right) = 0$ and asymptotic covariance matrix $\mathrm{Cov}\left(\hat{\beta}_{\mathrm{ML}}\right) = (X'\widehat{W}_{\mathrm{NB}}X)^{-1}$, Hence the asymptotic trace mean-squared error (TMSE) based on the asymptotic covariance matrix equals

$$\mathrm{TMSE}\left(\hat{\beta}_{\mathrm{ML}}\right) = \mathrm{E}(L^2_{\mathrm{ML}}) = \mathrm{E}(\hat{\beta}_{\mathrm{ML}} - \beta)'\left(\hat{\beta}_{\mathrm{ML}} - \beta\right),$$

$$= \mathrm{tr}\left[\left(X'\widehat{W}_{\mathrm{NB}}X\right)^{-1}\right] = \sum_{j=1}^{p}\frac{1}{\lambda_j}, \qquad (5)$$

where λ_j is the j^{th} eigenvalue of $\left(X'\widehat{W}_{\mathrm{NB}}X\right)$.

3. Negative Binomial Ridge Estimator

The ridge regression method proposed by Hoerl and Kennard (1970a, b) is well known as an efficient remedial measure in the presence of multicollinearity. The idea behind the ridge regression is that adding a small positive number k to the diagonal entries of the design matrix to decreases the variance; so that, one can obtain stable estimated coefficients.

When the explanatory variables are highly correlated the weighted matrix of cross-products $(X'\widehat{W}_{\mathrm{NB}}X)$ is ill-conditioned which leads to instability and high variance

of the ML estimator. In that situation, it is very hard to interpret the estimated parameters since the vector of estimated coefficients is on average too long. Ridge regression was extended to the class of GLMs for the logistic regression model by Schaefer et al. (1984), and then extended to the Poisson model by Månsson and Shukur (2011).

Mansson (2011) proposed a negative binomial ridge (NBR) regression estimator as a robust option of estimating the parameters of the NB model in the presence of multicollinearity. The NBR estimator is defined as follows;

$$\hat{\beta}_{\text{k.NB}} = (X'\widehat{W}_{\text{NB}}X + kI)^{-1}(X'\widehat{W}_{\text{NB}}X)\hat{\beta}_{\text{ML.NB}}, \tag{6}$$

where $k > 0$. The TMSE of this NBR estimator equals:

$$\begin{aligned}\text{TMSE}(\hat{\beta}_{\text{k.NB}}) &= E(\hat{\beta}_{\text{k.NB}} - \beta)' E(\hat{\beta}_{\text{k.NB}} - \beta),\\ &= \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j+k)^2} + \beta' k^2 (X'\widehat{W}_{\text{NB}}X + kI)^{-2}\beta,\\ &= \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j+k)^2} + k^2 \sum_{j=1}^{p} \frac{\alpha_j^2}{(\lambda_j+k)^2},\end{aligned} \tag{7}$$

whereα_j is defined as the j^{th} element of $\gamma\beta$ and γ is the eigenvector defined such that $(X'\widehat{W}_{\text{NB}}X) = (\gamma'\Lambda_{\text{NB}}\gamma)$, where Λ_{NB} equals diag(λ_j). By differentiating equation (7) with respect to k, setting to zero, and solving for k, we'll get the optimal value of k, which is $k = 1/\alpha_j^2$.

To investigate the performance of the NBR estimator and the traditional ML approach Mansson (2011) compared the performance of the estimators by conducting a simulation study. The performance of the estimator was compared by the other well-known ones and judged by calculating the MSE and the percentage of times ML outperforms NBR estimator. The results from the simulation study showed that the MSE decreases when the sample size becomes larger and the MSE becomes inflated as the other factors increase. Based on the result from the simulation study he recommended using the ridge parameters $K5-K7$ to practitioners, see table A.1 in the Appendix.

The selection of ridge parameter in biased estimators has always been an important issue, several methods of estimating the ridge parameter are proposed in the literature, table A.1 in Appendix, summarizes the different formula of ridge parameters used in the area of generalized linear models, corresponding to each author. For this study, we rely on $K1 = \frac{\hat{\sigma}^2}{\hat{\alpha}^2_{max}}$, K3= $\frac{\hat{\sigma}^2}{\left(\prod_{j=1}^{p} \hat{\alpha}_j^2\right)^{\frac{1}{p}}}$, K5= $Max\left(\frac{1}{m_j}\right)$, and K10= $median(m_j)$, to investigate the performance of ridge estimator and the ML estimator through a real dataset.

4. Application

In this section, the implementation of the methodology is illustrated by a study applied to a medium-sized timber industry which manufactures laminated plastic plywood by Filho and Sant'Anna (2016). The study consisted of evaluating the effect of input variables over the number of defects found in produced plywoods. The quality of the plywood is related to some variables, as detailed by Demirkir et al. (2013), Azaman et al. (2013), and Fang et al. (2014).

As in Filho and Sant'Anna (2016), we are considering the number of defects per laminated plastic plywood area as the independent variable (y) and the following input variables: volumetric shrinkage (x_1), assembly time (x_2), wood density (x_3), and drying temperature (x_4). So, for each sample unity representing a big wooden plate with constant size, we have data of the number of imperfections accompanied by the input data of the four process variables described. All the analysis in this paper is made using R language.

Table 1 shows the descriptive statistics for all variables of the study, one can see from the values of Skewness and kurtosis that independent variables are normally distributed whereas the dependent variable deviates away from normality. We can also notice that the mean of the response variable is 14.28, and the variance is 1936.88, which is enough evidence for over-dispersion.

Table 1: Descriptive Statistics for the dataset

		Min	Max	Mean	S.D	Skewness	Kurtosis	Unit
Imperfections	$\boldsymbol{y}$	0	404	14.28	44.01	7.24	59.18	-
Shrinkage	$\boldsymbol{x_1}$	7.46	12.39	9.63	1.22	0.27	-0.595	%
Assembly time	$\boldsymbol{x_2}$	12.6	17.90	14.99	1.15	0.195	-0.571	Minutes
Wood density	$\boldsymbol{x_3}$	.50	.57	.54	.02	0.0032	-0.495	g/cm^3
Drying temperature	$\boldsymbol{x_4}$	86.9	155.7	124.4	14.68	-0.25	-0.264	$^{\circ}C$

The correlation matrix in Table 2 shows the bivariate correlation between all variables associated with two-tailed significant test in parentheses. It is possible to notice that all process variables are significantly correlated with the response variable and the input variables x_1 and x_2 are highly correlated to each other (0.98), and (0.93) between x_3 and x_4. Thus, we are facing a case with evidence that the count variable changes according to the collinear input variables.

Table 2: Correlation Matrix

	$\boldsymbol{y}$	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_3}$	$\boldsymbol{x_4}$
$\boldsymbol{y}$	1				
$\boldsymbol{x_1}$	0.45 (.000)	1			
$\boldsymbol{x_2}$	0.43 (.000)	**0.98** (.000)	1		
$\boldsymbol{x_3}$	0.36 (.000)	-0.02 (.834)	-0.05 (.609)	1	
$\boldsymbol{x_4}$	0.39 (.000)	0.05 (.631)	-0.04 (.688)	**0.93** (.000)	1

Since the dependent variable is "the number of defects per laminated plastic plywood area" which is a nonnegative count variable, so the chosen model will be the Poisson or the negative binomial model. But, in case of over-dispersion the NB model is more favorable than the Poisson model.

Table 3 shows the model estimates using ML method associated with the multicollinearity diagnosis through the variance inflation factor (VIF) for each input variable, we can see that two of the important variables are not significant, and the multicollinearity index (VIF > 10) has a very big values which is confirmed the presence of a high correlation between the input variables.

Table 3: Parameter Estimates of the Negative Binomial Model

	Estimate	**SE**	**z-value**	**P-value**	**VIF**
(Intercept)	-20.3728	3.634	-5.606	2.07E-08	-
x_1	-0.62014	0.32927	-1.883	0.0596	223.204
x_2	1.50647	0.32987	4.567	4.95E-06	195.595
x_3	-10.8897	13.04345	-0.835	0.4038	48.48
x_4	0.08861	0.01528	5.8	6.64E-09	56.078

As a remedy to this problem, we apply the method of ridge regression for the NB model, using the selected ridge parameters that mentioned above, and compare the results with the ML results. The results can be found in Table 4 that present values of the estimated coefficients, the standard errors and the significance of each variable for different values of ridge parameter. Also, the trace standard error (TSE) is computed and listed in table 4. We can observe that the sign of estimates for x_1, and x_2, changes through the different estimators due to the server multicollinearity between these variables, whereas, it is still the same for x_3, and x_4, for the different estimators.

We can also see that the RR estimates and the associated standard errors are smaller than the ML method, the most substantial decrease in the parameter estimates and in the standard error can be found for variable x_3 when the ridge parameter *K*5 is used. Moreover, the values of the standard errors for *K1* are very close to the standard errors of ML estimator, and the values of the standard errors for *K3* and *K10* are somehow close to each other, while the values of the standard errors for *K5* are completely different from other estimates. This indicates that the multicollinearity problem leads to an estimated

value of the parameter associated with the x_3 variable that is larger than what it should be and thus that the effect of increasing the number of defect in the plywood is exaggerated.

Table 4: Ridge Regression for the NB Model

K1 (TSE=14.3783)	**Estimate**	**SE**	**z-value**	**P-value**
Intercept	-20.59255	3.03603	-6.78272	< .00001
x_1	-0.59930	0.28131	-2.13039	.016586
x_2	1.47997	0.28268	5.235496	< .00001
x_3	-10.00444	10.76558	-0.9293	.176445
x_4	0.05913	0.01268	4.663249	< .00001
K3 (TSE=0.8835)				
Intercept	-13.89811	0.52881	-26.2819	< .00001
x_1	0.57284	0.09202	6.225168	< .00001
x_2	0.20178	0.09183	2.197321	.014
x_3	-3.81615	0.16858	-22.637	< .00001
x_4	0.04317	0.00226	19.10177	< .00001
K5 (TSE=0.1378)				
Intercept	-1.36689	0.03670	-37.245	< .00001
x_1	2.00445	0.04891	40.98242	< .00001
x_2	-1.46372	0.03936	-37.188	< .00001
x_3	-0.45194	0.01078	-41.9239	< .00001
x_4	0.02158	0.00209	10.32536	< .00001
K10 (TSE=1.28038)				
Intercept	-18.31904	0.70794	-25.8765	< .00001
x_1	-0.03977	0.11269	-0.35292	.362419
x_2	0.87524	0.11619	7.532834	< .00001
x_3	-5.06465	0.34118	-14.8445	< .00001
x_4	0.04963	0.00238	20.85294	< .00001

It is worth noting that the significance of the model estimate is a good indicator of the efficiency of the estimation method. Therefore, by comparing the ML method with the ridge regression method, we can reach to the same conclusion of the superiority of the ridge regression method over the ML method, especially when the ridge parameter *K5* is applied.

5. Conclusion and Remarks

In this paper, the RR method of the NB regression model is presented as an efficient option to ML method when the explanatory variables are highly inter-correlated. Some ridge parameters are used to investigate the performance of the adapted methodology in this study. A real dataset is applied to evaluate the performance of the selected ridge estimators over the ML estimator. Results of application indicated that the ridge estimators outperform the ML estimator especially when using the ridge parameter *K5*. Therefore, we recommend using RR method with ridge parameter *K5*, when there are high inter-correlations among the independent variables.

There are some advantages of this paper over the work of Filho and Sant'Anna (2016), as follows;

1. The authors used the Poisson model in their work. But, because of the over-dispersion of the response variable, the NB model is more appropriate to fit the data than the Poisson model.
2. They adapted the PC regression method; PC regression method transforms the origin data by reducing the dimensions. So, it is too hard to interpret the newly variables or factors resulting by factor analysis. On the contrary, the ridge regression method retains the origin variables; therefore it is too easy to interpret the estimated model.
3. For future works, we can apply the Liu-type estimator proposed by Liu (2003), and extended to the NB model by Asar (2018), in order to get more efficient results.

Appendix

Table A.1: Ridge Parameters from the Literature

S.N	Author	Formulae	Such that
1)	Hoerl and Kennard (1970a,b)	$K1 = \hat{k}_{\mathrm{HK1}} = \frac{\hat{\sigma}^2}{\hat{\alpha}^2_{\max}}.$	Where $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2}{n-p-1}$, $\hat{\alpha}^2_{\max} =$ the maximum element of $\hat{\alpha}^2_{\mathrm{j}}$.
2)	Schaeffer et al (1984)	$K2 = \hat{k}_{\mathrm{SRW}} = \frac{1}{\hat{\alpha}^2_{max}}.$	
3)	Kibria (2003)	$K3 = \hat{k}_{\mathrm{GM}} = \frac{\hat{\sigma}^2}{\left(\prod_{j=1}^{p}\hat{\alpha}_j^2\right)^{\frac{1}{p}}},$ $K4 = \hat{k}_{\mathrm{MED}} = \mathrm{Median}\,\{m_j^2\}.$	Where $m_j^2 = \sqrt{\frac{\hat{\sigma}^2}{\hat{\alpha}_j^2}}.$
4)	Alkhamisi et al. (2006)	$K = \hat{k}^{\mathrm{KS}}_{\max} = \mathrm{Max}\,\{s_j\}$	Where $s_j = \frac{t_j\hat{\sigma}^2}{(n-p)\hat{\sigma}^2 + t_j\hat{\alpha}_j^2}$, and t_j is the eigenvalues of the $(X'X)$ matrix.
5)	Muniz and Kibria (2009)	$K5 = \hat{k}_{\mathrm{KM2}} = \mathrm{Max}\left(\frac{1}{m_j}\right),$ $K6 = \hat{k}_{\mathrm{KM4}} = (\prod_{j=1}^{p}\frac{1}{m_j})^{\frac{1}{p}},$ $K7 = \hat{k}_{\mathrm{KM6}} = \mathrm{median}\left(\frac{1}{m_j}\right).$	
6)	Kibria et al (2012)	$K8 = \max(m_j),$ $K9 = (\prod_{j=1}^{p} m_j)^{\frac{1}{p}},$ $K10 = \mathrm{median}(m_j),$ $K11 = \mathrm{Max}\left(\frac{1}{q_j}\right),$ $K12 = Max(q_j),$ $K13 = (\prod_{j=1}^{p}\frac{1}{q_j})^{\frac{1}{p}},$ $K14 = (\prod_{j=1}^{p} q_j)^{\frac{1}{p}},$ $K15 = \mathrm{median}\left(\frac{1}{q_j}\right),$ $K16 = \mathrm{median}(q_j).$	Where$\{ q_j = \frac{\lambda_{max}}{[(n-p)\hat{\sigma}^2+\lambda_{max}\hat{\alpha}_j^2]}$, λ_{max} = the maximum eigenvalue of $(X'\widehat{W}X)\}$.

References 0

Alkhamisi, M., Khalaf, G., & Shukur, G. (2006). Some modifications for choosing ridge parameters. *Communications in Statistics—Theory and Methods*, *35*(11), 2005-2020.

Asar, Y. (2018). Liu-type Negative Binomial Regression: a comparison of recent estimators and applications. In: Trends and perspectives in linear statistical inference. Springer, Cham, pp 23–39.

Azaman, M. D., Sapuan, S. M., Sulaiman, S., Zainudin, E. S., & Khalina, A. (2013). Shrinkages and warpage in the processability of wood-filled polypropylene composite thin-walled parts formed by injection molding. *Materials and Design 52*:1018–1026.

Cameron, A. C., & Trivedi, P. K. (1986). Econometric models based on count data: Comparisons and applications of some estimators. *Journal of Applied Econometrics* **1**: 29–53.

Demirkir, C., Özsahin, S., Aydin, I., & Colakoglu, G. (2013). Optimization of some panel manufacturing parameters for the best bonding strength of plywood. *International Journal of Adhesion & Adhesives 46:14–20.*

Fang, L., Chang, L., Guo, W. J., Chen, Y., & Wang, Z. (2014). Influence of silane surface modification of veneer on interfacial adhesion of wood–plastic plywood. *Journal of Applied Surface Science 288*:682–689.

Filho, D. M., & Sant'Anna, A. M. (2016). Principal component regression-based control charts for monitoring count data. *International Journal of Advanced Manufacturing Technology 85:1565–1574.*

Gujarati, D. N. (2009). *Basic Econometrics.* 5th Edition. New York: McGraw-Hill.

Hilbe, J. M. (2011). *Negative binomial regression*. 2nd Edition. Cambridge University Press.

Hoerl, A. E., & Kennard, R.W. (1970a). Ridge regression: biased estimation for non-orthogonal Problems. *Technometrics* 12, 55–67.

Hoerl, A. E., & Kennard, R.W. (1970b). Ridge regression: application to non-orthogonal problems. *Technometrics* 12, 69–82.

Kibria, B. G. (2003). Performance of some new ridge regression estimators. *Communications in Statistics-Simulation and Computation*, *32*(2), 419-435.

Kibria, B. G., Mansson, K., & Shukur, G. (2012). Performance of some logistic ridge regression estimators. *Computational Economics* 40(4):401–414.

Liu, K. (2003). Using Liu-type estimator to combat collinearity. *Communications in Statistics-Theory and Methods*, *32*(5), 1009-1020.

Månsson, K. (2011). On ridge estimators for the negative binomial regression model. *Economic Modelling*, *29*(2), 178-184.

Månsson, K., & Shukur, G. (2011). A Poisson ridge regression estimator. *Economic Modelling*, *28*(4), 1475-1481.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis*. 5th Edition. John Wiley & Sons.

Muniz, G., & Kibria, B. G. (2009). On some ridge regression estimators: An empirical comparison. *Communications in Statistics—Simulation and Computation*, *38*(3), 621-630.

McCullagh, P., & Nelder, J. A. (1989). Generalized linear models. 2nd Edition. Chapman & Hall, London.

Schaefer, R. L., Roi, L. D., & Wolfe, R. A. (1984). A ridge logistic estimator. *Communications in Statistics-Theory and Methods*, *13*(1), 99-113.

A Modification on the Weighting Scheme of Yitzhaki, Used for the OLS Test of Normality

Ahmed,A.E[1] *Osama, I.M.A*[2]

Abstract

Yitzhaki (2012) showed that the OLS estimator is a weighted average of the slopes defined by adjacent observations. The weights depend only on the distribution of the independent variable. In this article, the relationship between the distribution of the independent variable when it is uniformly distributed and the weighting scheme of the OLS estimator will be investigated. A modification on the yitzhaki (2012) result when the independent variable is uniformly distributed was introduced.

Keywords: Normality, Regression weights, Ordinary least squares, Monte Carlo Simulation.

Introduction

Statistical methods are based on various underlying assumptions. One common assumption is that a random variable is normally distributed or the errors ε_i are approximately normal.one of the normality tests used is the OLS test by shalit (2012).this test depends on the regression weights. These weights depend only on the distribution of the independent variable. Equal weights can only be obtained if, and only if the independent variable is normally distributed.

Yitzhaki (2012) proved that the regression weights are calculated using the following formula:

$$w(x) = \frac{1}{\sigma_X^2}\left[\mu_X F_X(x) - \Theta_X(x)\right] \qquad (1)$$

Where, $$\Theta_X(x) = \int_{-\infty}^{x} t f_X(t)dt = F_X(x)\, E\{X \mid X \le x\}$$

He investigated the relationship between the distribution of the independent variable and the weighting scheme of the OLS estimator assuming that the independent variable is a random variable.

[1] **Prof. of Statistics, ISSR, Cairo University**

[2] **Assist.Lec. Faculty of Commerce – Zagazig University**

When the independent variable is uniformly distributed and by applying eq.(1), the weighting function was:

$$w(x) = \frac{(x-a)(b-x)}{2(b-a)} \qquad (2)$$

Where a & b are the parameters of the uniform distribution function

A modification on his result is presented in the following lemma:

Lemma: if X is distributed as $Unif(a,b)$ by applying equation (1), the weight attached to the slope at point x has the form:

$$w(x) = \frac{6}{(b-a)^2}\left[\frac{(x-a)(b-x)}{(b-a)}\right]$$

Proof

Since,

$$F_X(x) = \frac{x-a}{b-a} \quad \& \quad , \quad \sigma_X^2 = \frac{(b-a)^2}{12} \qquad \mu = \frac{a+b}{2}$$

By using eq. (1), the weighting function will be,

$$w(x) = \frac{12}{(b-a)^2}\left[(\frac{a+b}{2})(\frac{x-a}{b-a}) - \Theta_X(x)\right]$$

$$\text{where} \quad \Theta_X(x) = \int_a^x t f_X(t)\,dt = \int_a^x \frac{t}{b-a}\,dt$$

$$= \frac{(x^2 - a^2)}{2(b-a)}$$

Therefore,

$$w(x) = \frac{12}{(b-a)^2}\left[\frac{(a+b)(x-a)(x-a)(x+a)}{2(b-a)}\right]$$

After some simplifications, the result will be obtained.

2. Simulation Study

The power of statistical tests should always include the empirical type I error rate of the test. The probability of this type I error rate of the test should be bounded upwards by the chosen level of significance; otherwise the test cannot be used for the given purpose. On the other hand, a test with a type I error rate far smaller or greater than the chosen α is an indicative of a test with low and high power respectively.

From tables (A1 and B1), it can be concluded that all tests are with acceptable type I error rates with all simulated type I error rates that are specified around the

level of significance except the JB test with all rates under 5% and tend to under reject at $\alpha = 0.05$, whereas when $\alpha = 0.10$, all tests have simulated type I error rates around the specified level of significance except the JB and RJB tests with rates lower than 10% and tend to under-reject the alternative hypothesis. The critical values of OLS test have been computed by generating 10,000 random samples from the standard normal distribution for different sample sizes and for different levels of significance as shown in table (A2).

Monte Carlo procedures were used to evaluate the power of CSQ, LF, AD, JB, SK, KU,B, RJB, SW, SF and OLS test statistics if a random sample of *(n)* independent observations come from a population with a normal $N\left(\mu,\sigma^2\right)$ distribution. The levels of significance (*α)* considered were 5% and 10%. 10,000 sample each of size *n* = 10, 15, 20, 30, 40, 100, 200 and 300 are generated from each of the given alternative distributions. The alternative distributions are. Weibull (3,1), Weibull (4,2), and Exp (8) which are asymmetric short-tailed. An asymmetric long – tailed distributions considered were Exp (4), Exp (10) , Weibull (2,2). All simulations were computed in R (version x64 3.2.2) except the values of the skewness and kurtosis of the alternative distributions used in the study were computed using MATHCAD (version14, 2007).

2.1 Level of significance α = 0.05

For Weibull (3,1) , the CSQ had the highest power at n = 10 and 15 while the KU test was the most powerful test for moderate sample sizes and the SW test was the most powerful one at n = 200 and 300 followed by the KU and the OLS tests respectively. For Weibull (4, 2), the CSQ was the most powerful one for small and moderate samples $\left(n \le 40\right)$ while the KU test was the most powerful one for large sample sizes. For the Exp (8), the SW was the most powerful test achieving 83.21% at n = 20 followed by the SF test with a power of 79.36% at this sample size. The B test was the least powerful test with a power of 58.87% at n = 100. It can be notice that when the size of the parameter of exponential distribution increase this tends to a slightly increase in the power of detecting non-normality. The B test was the least powerful test for these cases of the exponential distribution.

The OLS test had high power especially for small and moderate sample sizes. For Weibull (2,2) as an alternative distribution that is asymmetric with long tails, the SW was the most powerful test for all sample sizes followed by the SF and OLS tests respectively.

2.2 Level of significance α = 0.10

For Weibull (3,1), the CSQ, the KU and SW tests were the most powerful tests for small, moderate and large sample sizes respectively. While for Weibull (4, 2), the Pearson χ^2 and KU tests were the most powerful tests.

For Exp (8), the SW was the most powerful test for all samples followed by its modified form the SF test. With Exp (4) and Exp (10) as alternative distributions, the SW was the most powerful test followed by the SF and AD tests respectively. The B test was the least powerful, it had not reached the power of 100% even when n = 300 like the other tests. In situation where the Weibull (2, 2) is the alternative distribution that is asymmetric with long tails. The SW was the most powerful test at all sample sizes followed by the SF and OLS tests respectively.

3. Conclusion

(a) When α = 0.05

For asymmetric short-tailed, the SW was the most powerful test followed by the KU test against weibull distribution. The CSQ and the LF were the least powerful tests respectively. In case of asymmetric long-tailed distributions, the SW test was the most powerful one followed by its modified form the SF test while the B and KU tests were the least powerful ones respectively.

(b) When α = 0.10

For asymmetric short-tailed distributions, the SW test was the most powerful test followed by its modified form the SF test. While the CSQ and the LF tests were the least powerful ones respectively against the weibull distribution and the B and the KU against other tests. In case of asymmetric long-tailed, the SW and SF tests were the most powerful tests while the B and the KU tests were the least powerful tests respectively. In general, the OLS test had very high powers against most of alternative distributions especially for moderate and large samples

Table A1: Type I error rate at 5% significance level.

Normal (0, 1) – Skewness = 0, Kurtosis = 0

N	CSQ*	LF*	AD*	JB	SK Z_s *	KU Z_k *	B*	RJB*	SW*	SF*	OLS*
10	0.0608	0.0502	0.0514	0.0072	0.0517	0.0414	0.0408	0.0535	0.0500	0.0572	0.0478
15	0.0514	0.0483	0.0498	0.0163	0.0495	0.0390	0.0442	0.0592	0.0483	0.0516	0.0529
20	0.0493	0.0504	0.0499	0.0247	0.0520	0.0456	0.0472	0.0616	0.0501	0.0522	0.0498
30	0.0510	0.0469	0.0471	0.0323	0.0502	0.0505	0.0506	0.0633	0.0491	0.0521	0.0502
40	0.0628	0.0497	0.0523	0.0398	0.0534	0.0553	0.0499	0.0647	0.0513	0.0535	0.0504
100	0.0521	0.0517	0.0488	0.0415	0.0499	0.0517	0.0492	0.0559	0.0477	0.0500	0.0488
200	0.0536	0.0533	0.0508	0.0451	0.0504	0.0520	0.0510	0.0538	0.0517	0.0531	0.0512
300	0.0561	0.0492	0.052	0.0417	0.046	0.0503	0.0527	0.0452	0.0466	0.0466	0.0502

*Tests with acceptable type I error rates

Table A2: Critical Values for the OLS test obtained using 10,000 repetitions

Sample size (n)	α = 0.10	α = 0.05	α = 0.01
10	0.3640148	0.4154453	0.4983930
15	0.3084606	0.3478193	0.4223182
20	0.2749583	0.3101777	0.3790401
30	0.2288422	0.2561531	0.3129631
40	0.2033726	0.2289866	0.2776036
100	0.1299936	0.1465679	0.1785408
200	0.09399661	0.10564436	0.12765599
300	0.07661780	0.08523994	0.10534180

Table A3: Power for asymmetric short-tailed distributions at 5% significance level.

(a) Weibull (3, 1) – Skewness = 0.168, Kurtosis = 2.729

N	CSQ	LF	AD	JB	SK Z_s	KU Z_k	B	RJB	SW	SF	OLS
10	0.0676*	0.0468	0.0485	0.0076	0.0437	0.0349	0.0376	0.0461	0.0451	0.0476	0.0434
15	0.0473*	0.0465	0.0469	0.0116	0.0358	0.0425	0.0472	0.0420	0.0457	0.0417	0.0455
20	0.0457	0.0461	0.0414	0.0149	0.0328	0.0469*	0.0427	0.0356	0.0399	0.0361	0.0395
30	0.0524	0.0437	0.0483	0.0161	0.0359	0.0564*	0.0491	0.0347	0.0491	0.0400	0.0436
40	0.0583	0.0505	0.0493	0.0192	0.0377	0.0625*	0.0486	0.0320	0.0484	0.0375	0.0443
100	0.0589	0.0660	0.0707	0.0288	0.0609	0.0961*	0.0646	0.0335	0.0828	0.0563	0.0730
200	0.0680	0.0908	0.1152	0.0561	0.1128	0.1600	0.1082	0.0475	0.1607*	0.1090	0.1151
300	0.0839	0.1168	0.1687	0.1246	0.1667	0.2138	0.1430	0.0913	0.2727*	0.1823	0.1870

(b) Weibull (4, 2) – Skewness = -0.087, Kurtosis = 2.748

N	CSQ	LF	AD	JB	SK Z_s	KU Z_k	B	RJB	SW	SF	OLS
10	0.0606*	0.0453	0.0432	0.0054	0.0389	0.0347	0.0416	0.0398	0.0409	0.0419	0.0388
15	0.0516*	0.0473	0.0455	0.0104	0.0363	0.0368	0.0401	0.0432	0.0438	0.0425	0.0444
20	0.0466*	0.0438	0.0448	0.0127	0.0347	0.0394	0.0427	0.0387	0.0413	0.0390	0.0387
30	0.0471*	0.0446	0.0423	0.0132	0.0304	0.0437	0.0420	0.0300	0.0397	0.0317	0.0424
40	0.0566*	0.0472	0.0457	0.0125	0.0269	0.0508	0.0446	0.0288	0.0402	0.0316	0.0375
100	0.0556	0.0571	0.0614	0.0096	0.0284	0.0706*	0.0527	0.0146	0.0486	0.0297	0.0492
200	0.0655	0.0772	0.0885	0.0169	0.0397	0.1070*	0.0740	0.0141	0.0786	0.0467	0.0734
300	0.0745	0.0900	0.1149	0.0333	0.0517	0.1415*	0.0903	0.0200	0.1100	0.0638	0.1014

(c) Exp (8) – Skewness = 2, Kurtosis = 1.995

N	CSQ	LF	AD	JB	SK Z_s	KU Z_k	B	RJB	SW	SF	OLS
10	0.3933	0.2937	0.4072	0.1447	0.3709	0.2202	0.1101	0.3150	0.4419*	0.4290	0.3744
15	0.4891	0.4438	0.6224	0.3258	0.5611	0.2997	0.1643	0.4817	0.6768*	0.6468	0.6013
20	0.6464	0.5736	0.7762	0.4798	0.6998	0.3681	0.2008	0.5939	0.8321*	0.7936	0.7464
30	0.8459	0.7705	0.9289	0.7203	0.8748	0.4805	0.2743	0.7669	0.9655*	0.9456	0.9164
40	0.9505	0.9066	0.9836	0.8829	0.9610	0.5783	0.3217	0.8930	0.9947*	0.9899	0.9787
100	1.0000	1.0000	1.0000	1.0000	1.0000	0.8867	0.5887	0.9998	1.0000	1.0000	1.0000
200	1.0000	1.0000	1.0000	1.0000	1.0000	0.9893	0.8191	1.0000	1.0000	1.0000	1.0000
300	1.0000	1.0000	1.0000	1.0000	1.0000	0.9994	0.9167	1.0000	1.0000	1.0000	1.0000

Table A4: Power for asymmetric long-tailed distributions at 5% significance level.

(a) Exp (4) – Skewness = 2, Kurtosis = 9

N	CSQ	LF	AD	JB	SK Z_s	KU Z_k	B	RJB	SW	SF	OLS
10	0.3953	0.3014	0.4143	0.1468	0.3663	0.2214	0.1094	0.3114	0.4453*	0.4314	0.3812
15	0.4991	0.4561	0.6344	0.3235	0.5606	0.2970	0.1655	0.4785	0.6904*	0.6536	0.6112
20	0.6548	0.5738	0.7749	0.4858	0.7067	0.3666	0.2009	0.5951	0.8363*	0.7991	0.7463
30	0.8490	0.7731	0.9313	0.7217	0.8812	0.4769	0.2697	0.7752	0.9657*	0.9462	0.9187
40	0.9542	0.8973	0.9840	0.8789	0.9615	0.5821	0.3364	0.8879	0.9946*	0.9901	0.9762
100	1.0000	0.9999	1.0000	1.0000	1.0000	0.8801	0.5800	0.9998	1.0000	1.0000	1.0000
200	1.0000	1.0000	1.0000	1.0000	1.0000	0.9905	0.8134	1.0000	1.0000	1.0000	1.0000
300	1.0000	1.0000	1.0000	1.0000	1.0000	0.9987	0.9137	1.0000	1.0000	1.0000	1.0000

(b) Exp (10) – Skewness = 2, Kurtosis = 4.274

N	CSQ	LF	AD	JB	SK Z_s	KU Z_k	B	RJB	SW	SF	OLS
10	0.3975	0.2995	0.4171	0.1472	0.3754	0.2247	0.1121	0.3189	0.4528*	0.4424	0.3815
15	0.4948	0.4442	0.6238	0.3256	0.5588	0.3002	0.1665	0.4748	0.6806*	0.6471	0.6016
20	0.6510	0.5741	0.7735	0.4795	0.6968	0.3667	0.2033	0.5916	0.8358*	0.7967	0.7425
30	0.8517	0.7799	0.9345	0.7329	0.8874	0.4879	0.2720	0.7834	0.9677*	0.9533	0.9259
40	0.9516	0.8982	0.9837	0.8770	0.9606	0.5705	0.3250	0.8860	0.9958*	0.9905	0.9758
100	1.0000	0.9999	1.0000	0.9999	1.0000	0.8867	0.5824	0.9999	1.0000	1.0000	1.0000
200	1.0000	1.0000	1.0000	1.0000	1.0000	0.9891	0.8107	1.0000	1.0000	1.0000	1.0000
300	1.0000	1.0000	1.0000	1.0000	1.0000	0.9992	0.9176	1.0000	1.0000	1.0000	1.0000

(c) Weibull (2, 2) – Skewness = 0.631, Kurtosis = 3.245

N	CSQ	LF	AD	JB	SK Z_s	KU Z_k	B	RJB	SW	SF	OLS
10	0.0869	0.0755	0.0823	0.0200	0.0873	0.0627	0.0507	0.0789	0.0871	0.0904*	0.0805
15	0.0752	0.0845	0.1051	0.0432	0.1114	0.0728	0.0503	0.0986	0.1165*	0.1135	0.1136
20	0.0773	0.1061	0.1353	0.0701	0.1486	0.0878	0.0608	0.1264	0.1624*	0.1514	0.1458
30	0.0991	0.1369	0.1919	0.1174	0.2154	0.1119	0.0650	0.1645	0.2362*	0.2091	0.2240
40	0.1228	0.1631	0.2422	0.1597	0.2849	0.1275	0.0730	0.2005	0.3169*	0.2752	0.2607
100	0.2537	0.3899	0.6035	0.4973	0.6917	0.1509	0.0838	0.4713	0.7861*	0.7141	0.6740
200	0.5566	0.7171	0.9353	0.9397	0.9601	0.1701	0.0894	0.9074	0.9942*	0.9853	0.9521
300	0.7994	0.8818	0.9945	0.9967	0.9957	0.1969	0.1034	0.9953	1.0000*	0.9999	0.9965

Table B1: Type I error rate at 10% significance level.

***Normal* (0, 1) – Skewness = 0, Kurtosis = 0**

N	CSQ *	LF*	AD*	JB	SK Z_s *	KU Z_k *	B*	RJB	SW*	SF*	OLS*
10	0.1171	0.0980	0.1047	0.0168	0.1033	0.0877	0.0951	0.0725	0.0999	0.1104	0.1020
15	0.1150	0.0971	0.1013	0.0307	0.1050	0.0956	0.1018	0.0805	0.1015	0.1050	0.1045
20	0.1246	0.1028	0.1041	0.0373	0.1031	0.0949	0.0982	0.0851	0.1043	0.1078	0.1013
30	0.1214	0.1043	0.1046	0.0463	0.0983	0.0999	0.1075	0.0844	0.1033	0.1036	0.1062
40	0.1093	0.0933	0.0995	0.0505	0.0980	0.0971	0.0952	0.0816	0.1009	0.0998	0.0919
100	0.1009	0.1082	0.1006	0.0600	0.0964	0.0968	0.1018	0.0801	0.0950	0.0966	0.1030
200	0.1072	0.1044	0.1009	0.0751	0.0986	0.1009	0.0958	0.0802	0.0963	0.0995	0.0993
300	0.1053	0.1046	0.1012	0.0811	0.1007	0.0991	0.0999	0.0795	0.0996	0.1015	0.1069

*Tests with acceptable type I error rates

Table B2: Power for asymmetric short-tailed distributions at 10 % significance level.

(a) Weibull (3,1) – Skewness = 0.168, Kurtosis = 2.729

N	CSQ	LF	AD	JB	SK Z_s	KU Z_k	B	RJB	SW	SF	OLS
10	0.1072*	0.0913	0.0961	0.0137	0.0834	0.0865	0.1037	0.0558	0.0939	0.0931	0.0921
15	0.1108*	0.0883	0.0900	0.0196	0.0813	0.0894	0.0984	0.0551	0.0893	0.0829	0.0863
20	0.1280*	0.0969	0.1004	0.0237	0.0817	0.0988	0.1067	0.0570	0.0994	0.0885	0.0896
30	0.1147*	0.0952	0.1023	0.0297	0.0839	0.1077	0.1008	0.0508	0.1023	0.0856	0.0955
40	0.1134	0.1006	0.1058	0.0321	0.0850	0.1157*	0.1113	0.0506	0.1105	0.0864	0.0927
100	0.1141	0.1274	0.1406	0.0560	0.1255	0.1727*	0.1436	0.0567	0.1607	0.1141	0.1406
200	0.1273	0.1621	0.1977	0.1392	0.2019	0.2458	0.1842	0.1076	0.2807*	0.1949	0.2033
300	0.1468	0.2025	0.2750	0.2697	0.2833	0.3087	0.2349	0.2057	0.4196*	0.3068	0.2916

(b) Weibull (4,2) – Skewness = -0.087, Kurtosis = 2.748

N	CSQ	LF	AD	JB	SK Z_s	KU Z_k	B	RJB	SW	SF	OLS
10	0.1063*	0.0924	0.0983	0.0137	0.0912	0.0833	0.0968	0.0642	0.0936	0.0991	0.0984
15	0.1077*	0.0901	0.0915	0.0200	0.0772	0.0810	0.0995	0.0591	0.0903	0.0866	0.0882
20	0.1218*	0.0923	0.0907	0.0236	0.0773	0.0855	0.0983	0.0548	0.0896	0.0817	0.0846
30	0.1143*	0.0992	0.0977	0.0240	0.0717	0.0872	0.0992	0.0480	0.0924	0.0790	0.0895
40	0.1150*	0.1043	0.1013	0.0221	0.0713	0.0926	0.0984	0.0442	0.0930	0.0743	0.0865
100	0.1116	0.1211	0.1237	0.0262	0.0710	0.1251*	0.1164	0.0322	0.1103	0.0733	0.1122
200	0.1200	0.1412	0.1503	0.0594	0.0882	0.1722*	0.1407	0.0420	0.1447	0.0938	0.1370
300	0.1343	0.1686	0.1965	0.1104	0.1120	0.2328*	0.1783	0.0720	0.1963	0.1280	0.1776

(c) Exp (8) – Skewness = 2, Kurtosis = 1.995

N	CSQ	LF	AD	JB	SK Z_s	KU Z_k	B	RJB	SW	SF	OLS
10	0.4724	0.4174	0.5385	0.1954	0.4911	0.3048	0.1844	0.3585	0.5674*	0.5555	0.5191
15	0.5645	0.5698	0.7267	0.3929	0.6771	0.3835	0.2281	0.5190	0.7785*	0.7473	0.7098
20	0.7502	0.7030	0.8599	0.5719	0.8101	0.4636	0.2765	0.6516	0.9073*	0.8807	0.8415
30	0.8963	0.8721	0.9663	0.8048	0.9390	0.5711	0.3398	0.8242	0.9862*	0.9751	0.9601
40	0.9660	0.9480	0.9934	0.9311	0.9827	0.6671	0.4024	0.9209	0.9982*	0.9967	0.9910
100	1.0000	1.0000	1.0000	1.0000	1.0000	0.9219	0.6418	1.0000	1.0000	1.0000	1.0000
200	1.0000	1.0000	1.0000	1.0000	1.0000	0.9936	0.8570	1.0000	1.0000	1.0000	1.0000
300	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9386	1.0000	1.0000	1.0000	1.0000

Table B3: Power for asymmetric long-tailed distributions at 10 % significance level.

(a) Exp (4) – Skewness = 2, Kurtosis = 9

N	CSQ	LF	AD	JB	SK Z_s	KU Z_k	B	RJB	SW	SF	OLS
10	0.4634	0.4185	0.5366	0.2001	0.4922	0.3101	0.1807	0.3578	0.5695*	0.5563	0.5147
15	0.5658	0.5682	0.7265	0.3853	0.6738	0.3836	0.2313	0.5145	0.7814*	0.7451	0.7097
20	0.7503	0.6998	0.8531	0.5555	0.8063	0.4478	0.2717	0.6418	0.9050*	0.8737	0.8344
30	0.8986	0.8653	0.9658	0.8101	0.9453	0.5720	0.3479	0.8241	0.9867*	0.9773	0.9642
40	0.9673	0.9473	0.9928	0.9341	0.9821	0.6620	0.4078	0.9242	0.9986*	0.9962	0.9895
100	1.0000	1.0000	1.0000	1.0000	1.0000	0.9208	0.6411	1.0000	1.0000	1.0000	1.0000
200	1.0000	1.0000	1.0000	1.0000	1.0000	0.9946	0.8616	1.0000	1.0000	1.0000	1.0000
300	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9386	1.0000	1.0000	1.0000	1.0000

(b) Exp (10) – Skewness = 2, Kurtosis = 4.279

N	CSQ	LF	AD	JB	SK Z_s	KU Z_k	B	RJB	SW	SF	OLS
10	0.4668	0.4098	0.5337	0.1972	0.4804	0.3020	0.1837	0.3512	0.5646*	0.5495	0.5085
15	0.5652	0.5773	0.7306	0.3935	0.6766	0.3873	0.2347	0.5222	0.7850*	0.7517	0.7179
20	0.7448	0.6976	0.8578	0.5663	0.8155	0.4544	0.2693	0.6528	0.9077*	0.8805	0.8410
30	0.8996	0.8691	0.9687	0.8084	0.9433	0.5715	0.3490	0.8249	0.9865*	0.9778	0.9610
40	0.9678	0.9521	0.9932	0.9352	0.9824	0.6675	0.4016	0.9240	0.9986*	0.9971	0.9907
100	1.0000	1.0000	1.0000	1.0000	1.0000	0.9280	0.6546	1.0000	1.0000	1.0000	1.0000
200	1.0000	1.0000	1.0000	1.0000	1.0000	0.9943	0.8543	1.0000	1.0000	1.0000	1.0000
300	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9407	1.0000	1.0000	1.0000	1.0000

(c) Weibull (2,2) – Skewness = 0.631, Kurtosis = 3.245

N	CSQ	LF	AD	JB	SK Z_s	KU Z_k	B	RJB	SW	SF	OLS
10	0.1363	0.1296	0.1492	0.0279	0.1434	0.1141	0.1144	0.0912	0.1507*	0.1493	0.1465
15	0.1407	0.1498	0.1840	0.0659	0.1882	0.1381	0.1205	0.1287	0.2003*	0.1884	0.1914
20	0.1811	0.1787	0.2239	0.0968	0.2370	0.1541	0.1243	0.1501	0.2561*	0.2347	0.2359
30	0.1884	0.2201	0.2881	0.1635	0.3323	0.1781	0.1298	0.2079	0.3528*	0.3157	0.3211
40	0.2056	0.2718	0.3671	0.2201	0.4308	0.1908	0.1358	0.2556	0.4653*	0.4091	0.4054
100	0.3810	0.5509	0.7428	0.6678	0.8102	0.2386	0.1545	0.6070	0.8799*	0.8293	0.8008
200	0.6802	0.8269	0.9707	0.9795	0.9825	0.2574	0.1657	0.9666	0.9979*	0.9938	0.9808
300	0.8755	0.9416	0.9974	0.9995	0.9987	0.2830	0.1786	0.9991	1.0000	1.0000	0.9989

References

1. Anderson, T.W. and Darling, D.A. (1954). A Test of Goodness of Fit, JASA, 49, 765-769.
2. Anscombe, F.J. and Glynn, W.J. (1983). Distribution of Kurtosis Statistic b_2 for Normal Statistics. Biometrika, 70, 227–234
3. Bonett, D.G. and Seier, E. (2002). A Test of Normality with High Uniform Power. Computational Statistics & Data Analysis, 40 , 435 – 445.
4. Jarque, C.M. and Bera, A.K. (1980). Efficient test for normality, homoscedasticity and serial independence of regression residuals. Economics Letters, 6, 255–259.
5. Lilliefors, H.W. (1967). On the Kolmogorov–Smirnov test for normality with mean and variance unknown. JASA, 62, 399–402
6. Ramos, F.M, and Burgos, J.G. (2012). A Power Comparison of Various Tests of Univariate Normality on Ex- Gaussian Distributions. European Journal of Research Methods for the Behavioral and Social Sciences (Methodology), 1-13.
7. Razali N. and Wah Y. (2011). Power Comparison of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson Darling tests. Journal of Statistical Modeling and Analytics, 2 (1), 21-33.
8. Shalit, H. (2012). Using OLS to Test for Normality. Statistics and Probability Letters, 82 , 2050–2058.
9. Yitzhaki, Shlomo, (2012), "On Using Linear Regressions in Welfare Economics", Journal of Business and Economic Statistics, 14, 478-486.

Handling Mixed Missing Data with Application

Yasmin Mohamed Ibrahim[1] Mai Ahmed Mohsen [2]

Abstract:

Various techniques have been developed for dealing with missing values in data sets with homogeneous attributes (their independent attributes are all either continuous or discrete). However, these imputation algorithms cannot be directly applied to many real data sets, as survey data sets in general often consist of large numbers of variables which have mixed data types i.e. different measurement scales. Specific methods and modification in existing methods are found for dealing with such kind of data.

This paper reviews some methods for such kind of data and applies six imputation methods out of them. Assessing the performance of the six imputation methods which are MICE, MICE-CART, MICE-RF, MissForest, MissRanger and KNN is performed using a real dataset at 5 different missing rates. Results were assessed using different criteria.

Keywords:

MICE, MICE-CART, MICE-RF, MissForest, MissRanger KNN, missing at random, mixed data.

[1]Applied Statistics &Econometrics Department, Institute of Statistical Studies and Research, Cairo University

[2] Mathematics and Statistics Department, Sadat Academy for Management Sciences.

1. Introduction:

Surveys are mainly conducted to obtain valuable information on some criteria from a specified sample or population. But, the survey results often include non-response of the subjects under study for highly significant attributes **(Devi and Sivaraj, 2017)**. This missingness happens due to different reasons as; lack of response, machine error, history of data not registered, data collection not done properly, mistakes due to data entry operator and many other reasons. No matter what the reason is, incomplete data is often unavoidable pervasive problem faced by most applied researchers.

Treating methods for missing values varies widely, but most of these methods are based on three strategies: ignoring the missing data, imputing the missing data and using model-based procedures.

Ignoring the missing data is known as complete case analysis in which the data set is edited to ignore the missing data and consider only the non-missing data as well as omitting units with missing data from the study (listwise and pairwise).

Imputing the missing data on the other hand utilizes all collected information and allows the user to perform analysis of "complete" data, as the missing values are filled in with one (single imputation) or many (multiple imputation) “plausible” values. By single imputation, each missing observation is filled in by one imputed value, creating one "complete" data set. While in multiple imputation missing values are imputed from their posterior predictive distribution, this is done M times to generate M completed data sets, the purpose is to obtain estimates that better reflect the true variability and uncertainty in the data **(Hörnblad, 2013).**

Modelling techniques are generated by factorization of the likelihood according to the observation pattern and the missing pattern. Parameters can be estimated by iterative maximum likelihood procedures, one example for model-based procedure is the well-known EM algorithms.

Not only the amount of the missing data but also the mechanism of missingness, represent a critical issue a researcher must address before choosing an appropriate procedure to deal with missing data. Different missing mechanisms are addressed in the next section

2. Missing Data Mechanism

The missing data mechanism describes the relationship between the missingness of the data and the values of the variables in the data matrix, i.e. whether the missingness depends on the underlying values of the variables in the data set. Mechanisms that lead to missing data can be classified as: missing completely at random, missing at random, and non-ignorable missing.

2.1 Missing Completely at Random (MCAR):

The mechanism is said to be missing completely at random (MCAR) When missing items do not depend upon both observed values and missing values of Y; that is if

$$P(R|Y_{obs}, Y_{mis}, \Phi) = P(R|\Phi) \qquad 1.1$$

Where Φ refers to the parameters of the missing data mechanism, Y_{obs} is the observed values of Y, and Y_{mis} the missing values of Y. For example, a question in a survey was overlooked for certain respondent; that is MCAR. In practice this assumption is too restrictive.

2.2 Missing at Random (MAR)

Under the missing at random **(**MAR) assumption the missingness is allowed to depend on the observed data (Y_{obs},), but not directly on the missing data (Y_{mis}). The data is MAR if

$$P(R|Y_{obs}, Y_{mis}, \Phi) = P(R|Y_{obs}, \Phi), \qquad 1.2$$

For example, perhaps males are more likely to drop out than females in a smoking study because they do not want to participate in a smoking cessation program **(Rashid, 2011),** or if men are more likely to tell you their weight than women, so weight is MAR.

2.3 Not Missing at Random (NMAR)

If the missing data cannot be assumed to be neither MCAR nor MAR, then the data is said to be not missing at random (NMAR) or non-ignorable**.** That happens if the probability of missing data depends on the missing values themselves. That is, the probability of R depends on the missing values even after taking the observed values into account

$$P(R|Y_{obs}, Y_{mis}, \Phi) = P(R|Y_{obs}, Y_{mis}, \Phi) \qquad 1.3$$

This is a case where the people with the lowest education are missing on education or the sickest people are most likely to drop out of the study.

3. Handling Missing values in mixed data

The literature of missing values in mixed data can be classified to: 1) Using joint modeling technique for handling missing values in mixed data (**Little and Schluchter, 1985)**, **(Ganjali, 2003**), **(Bahrami et al., 2010) and** (**Amiri et al,**

2017). 2) Multiple imputation by chained equation for mixed data (**Raghunathan et al., 2001), (Buuren et al, 2006) and (Lee and Carlin, 2010**). 3) Neighborhood and distance metrics techniques **(Ayuyev et al., 2009) (Tarsitano and Falcone, 2011) (Liao et al., 2014) and (Sen et al., 2018)**. 4) Techniques that relies on principal component **(Ren, 2012) (Audigier et al., 2016).** 5) Techniques that relies on random forests **(Stekhoven and Buhlmann, 2011) (Doove et al., 2014) (Shah et al., 2014)** 6) Various other techniques as mixture kernel technique **(Dane and Thool, 2013) and** sequential regression fractional imputation procedure **(She, 2017).** In the next section we shall focus on the algorithms of K-Nearest Neighborhood (KNN), multiple imputation by chained equations (MICE) and imputation by random forests.

3.1 K Nearest Neighbor Methods for Mixed data

In K nearest neighbor (NN) each missing value is replaced by values obtained from observations (K donors) that are similar to the observation with the missing value, with respect to some observed characteristics. Four of the most popular techniques to deal with the presence of variables with different measurement scales using NN methods will be discussed below.

a. Dominant type approach: The simplest way is to divide the variables into types and conduct the analysis to the dominant type.

b. Converting one type of variable to another: Another approach is to convert one type of variable to another and then use a distance that is suitable for the selected type. A drawback of this approach is the use of a large number of binary variables that are highly interdependent. Alternatively, numerical variables could be categorized at a fixed level so that the new values can be treated using a categorical

distance function. A consequence is reducing the influence of the quantitative variables **(Tarsitano and Falcone, 2011)**.

c. Compromise distance: (Tarsitano and Falcone, 2011) suggest not to focus on the computation of a distance, but achieving a compromise distance using a combination of all the partial distances. Partial because each of them is linked to a specific type of variable. The separate distance matrices are combined as a weighted average, and the resulting global distance matrix is then used in the search for donors. The global distance may have the following form

$$\delta_{i,j} = \sum_{t=1}^{P} [h_{i,j}^{t} + \left(1 + h_{i,j}^{t}\right)\delta_{i,j}^{t}\,, \tag{3.1}$$

$$and\ \ h_{i,j}^{t} = \frac{\sum_{s=M_t-1}^{M_t} h_{s,i,j}}{m_t}\ \ , M_t = \sum_{s=1}^{t} m_s, M_0 = 0$$

Where $\delta_{i,j}^{t}$ is the t[th] partial distance between the records R_i and R_j. Usually the distances are scaled to vary in the unit interval between 0 and 1. Or it may be a Gower's distance

$$\text{Gower's distance} = \frac{\sum_{v=1}^{V} \delta_{ijv} d_{ijv}}{\sum_{v=1}^{V} \delta_{ijv}} \tag{3.2}$$

where d_{ijv} is the dissimilarity score between subject i and j for the v[th] variable and δ_{ijv} takes the value of 1 if the v[th] variable is available for both subject i and j and 0 otherwise. Depending on different types of variable, d_{ijv} is defined.

3.2 Multiple Imputations by Chained Equations (MICE)

The basic idea of MICE has been proposed by many researchers and is widely used under different names as: full conditional specification, stochastic relaxation, regression switching, sequential regressions, variable-by-variable imputation, ordered pseudo-Gibbs sampler, partially incompatible MCMC, and iterated univariate imputation (**Buuren and Groothuis, 2011).** The chained equations process goes through the following steps:

- Step 1:replace each missing value in the data with a simple imputation, these simple imputations can be thought of as "place holders".
- Step 2: the place holder for a target variable y (just one variable) is set back to missing.
- Step 3: the observed values from the target variable (y) is regressed on the other variables in the data (all or part of the variables in the dataset). These regression models should be the appropriate model (For example; binary logistic regression is used for dichotomous variables, polytomous logistic regression for ordinal variables, poisson regression for count variables, and ordinary least squares regression for continuous variables). It operates under the same assumptions that one would make when performing regression models outside the context of imputing missing data.
- Step 4: replace the missing values for "y" with predictions from the regression model. These predicted values are used along with the observed values as an independent variable in the regression models.
- Step 5: steps 2-4 are then repeated for each variable with missing data.
- Step 6: steps 2 through 4 are repeated for a given number of cycles or iterations, with the imputations being updated at each iteration. Cycles go until the

imputations have converged over the iterations in the sense of being stable. These cycles end with one imputed dataset.

- Step 7: the entire imputation process is repeated m times to generate multiple imputed datasets, where m = the desired number of multiple imputations.

MICE suffer from some limitations and obstacles among them; justification of the MICE procedure has rested on empirical studies rather than theoretical arguments. Also, the relation between the dependent and the independent variables could be nonlinear or of any complex form. Imputation can create impossible combinations as pregnant fathers or current weight of the dead **(Buuren and Groothuis, 2011).** Besides that, MICE is based on MAR assumption, so it is sensitive to work with data other than MAR, especially with larger fractions of missing data.

3.3. Random Forests Approach

Random forest is an extension of classification and regression trees, predictive models that recursively subdivide the data based on values of the predictor variables. They do not rely on distributional assumptions and can accommodate nonlinear relations and interactions **(Shah et. al., 2014)**. Random Forest (RF) approaches to imputation can be classified to three approaches; The proximity approach, on the fly approach and. MissForest approach. We shall focus on MissForest approach.

The MissForest approach works by recasting the missing data problem as a prediction problem. It starts by pre-imputed data; Then, sort the variables X_s, s = 1,...,p according to the amount of missing values starting with the lowest amount. For each variable X_s the missing values are imputed by first fitting a random forest then, predicting the missing values and so on growing a forest and iterating for improving the results. So, data is imputed by regressing each variable in turn against

all other variables and then predicting missing data for the dependent variable using the fitted forest **(Stekhoven and Buhlmann, 2011).** The stopping criterion is met as soon as the difference between the newly imputed data matrix and the previous one increases for the first time.

3.4 Multiple Imputation by Chained Equations Using Classification and Regression Trees, and Random Forests (MICE–CART and MICE- RF)

Because of the desired characteristics of CART and RF some missing data algorithms have recently been developed to incorporate CART and RF with the traditional imputation method. For instance, **(Doove et al., 2014)** proposed using CART and random forest for multiple imputation within the MICE framework. MICE-CART is based on the MICE algorithm but replaces the regression model with the CART algorithm. MICE-CART works as follows: 1. The missing values are initially imputed. 2. A tree is fitted on the first variable with at least one missing value, using the remaining variables as predictors. A member with a missing value on Y_1 is put down this tree and ends up in one of the leaves; use it to impute the missing value. 3. Repeat step 2 for every variable with missing value times creating one imputed dataset. 4. Repeat steps 1-3 m times, yielding m imputed sets.

While implying MI for RF goes as follows: 1. Draw k bootstrap samples from Y, restricted to members in Y with observed values. 2. Fit one tree on every bootstrap sample drawn in step 1. This results in k trees, where every tree has several leaves. Each leaf includes a subset of Y with observed values, which will be called donors. 3. For members with missing values in Y determine in which leaf they will end up according to the k trees fitted in step 2. 4. Take all donors from the K leaves ended up in step c together and randomly select one of the observed values from the donors.

Replace the originally missing values of Y with these imputation values. 5. Repeat step 2 to have performed it (number of iterations) times. 6. Repeat steps 1–3 m times, yielding m imputed sets. 7. This process was embedded into MICE and repeated to create multiple imputations. **(Shah et al., 2014)** also proposed using random forest for imputation using a somewhat different approach.

4. Application on Real Dataset

Imputation methods is applied on the National Health and Nutrition Examination Survey (NHANES) 2007/2008 data set. The National Health and Nutrition Examination Survey is a program of studies designed to assess the health and nutritional status of adults and children in the United States. NHANES is a major program of the National Center for Health Statistics (NCHS). The Weight History section of the sample person questionnaire provides personal interview data on several topics related to body weight, including; self-perception of weight, self-reported weight over the participant's lifetime, attempted weight loss during the past 12 months, and methods used to try to lose weight, and to keep from gaining weight…etc.

6546 respondents were interviewed, 3501Complete respondent cases were only selected that has no missingness or nonapplicable in the selected variables. The questionnaire consists of 99 variables, 30 of them are only selected. Of the selected variables 8 are binary, 7 are nominal (categorical), 6 are ordinal and 9 are numerical.

The performance of the different imputation methods will be assessed using two schemes:1. Evaluating the performance of the imputation values. 2. Evaluating the analyzed model after imputing the values using different imputation techniques.

4.1 Evaluating the Imputed Values

After imputing the missing values, the performance is assessed using the following two criteria: A. Normalized root mean squared error (NRMSE) for the continuous variables which is defined as:

$$\text{NRMSE} = \frac{\sqrt{\boldsymbol{mean}(x^{true} - x^{imp})^2}}{\boldsymbol{var} x^{true}} \qquad 4.1$$

B. The proportion of false classified entries (PFC) for categorical variables

$$\text{PFC} = \frac{\sum_{j \in f} \sum_{i=1}^{n} I X_{new}^{imp} \neq X_{old}^{imp}}{\#NA} \qquad 4.2$$

Where #NA is the number of missing values in the categorical variables. In both cases good performance leads to a value close to 0 and bad performance to a value around 1 **(Stekhoven and Buhlmann 2011).**

Since there is more than one imputed dataset for MICE, MICE-RF and MICE-CART. The NRMSE and the PFC will be the average of each imputed dataset**.**

4.2. Evaluating the Analyzed Model

Because multiple imputations involve creating multiple predictions for each missing value, the analyses of multiply imputed data take into account the uncertainty in the imputations and yield accurate standard errors. That's why one method for assessing the performance of the imputation method is through the analyzed model not just the imputed values, the used strategy will be in the following algorithm:1. Determine the analyzed model from the complete data 2. Create missingness in the independent variables but keep the dependent variable complete as it is. 3. Fit the model using the imputed data set and pool the results for multiple

imputed datasets.4. Compare the results of the model from the complete data and the model from the imputed values using the following criteria:

A. Coefficients of the Model

(RAAD) is the average of the absolute difference between the coefficients of the model from complete data (the full model) and the coefficient of the model from the imputed data (the imputed model) divided by the average of absolute coefficients of the complete model and multiplied by 100.

B. The Standard Error (SE)

the Relative Average Standard Error of the coefficient (RASE) will be calculated as

$$RASE = \frac{average\ SE\ (for\ \text{coefficients of}\ imputed\ data\ model\ )}{average\ SE(for\ \text{coefficients of}\ complete\ data\ model\ )} \times 100 \qquad 4.3$$

C. Coefficient of Determination (R^2)

The R-squared of the regression is the fraction of the variation in the dependent variable that is accounted for (or predicted by) the independent variables.

$$R^2=\frac{SSR.}{SS_{\text{total}}} \qquad 4.4$$

D. Mean Square Error (MSE)

Mean Square Error (MSE) is the average of the square of the errors. Error in this case means the difference between the observed values y_1, y_2,…y_n and the predicted ones $\widehat{y_1}, \widehat{y_2}$,... $\widehat{y_n}$

$$MSE=\frac{(yi-\hat{y}_i)^2}{n-1} = \frac{ss_{res}}{df} \qquad 4.5$$

To compare across different models, Relative Mean Square Error (RMSE) will be used; which is

$$\frac{MSE\ (for\ imputed\ data\ model\ )}{MSE(for\ complete\ data\ model\ )} \times 100 \qquad 4.6$$

5. Results and Discussion

For evaluating the imputed values; as shown in Table 4.1 MissForest has the lowest NRMSE across all the missing rates and KNN has the highest across all the missingness rates also, it is followed by MICE-RF which has the second order in high NRMSE across all the missing rates.

Table 4.1: Normalized Mean Square Error (NRMSE) for the National Health and Nutrition Examination Survey (NHANES) 2007/2008 Dataset

Missing Rate	Methods					
	MICE	**MICE-RF**	**MICE-CART**	**Miss-Forest**	**Miss-Ranger**	**KNN**
5%	0.21	0.25	0.22	**0.15** [a]	0.20	**0.26** [b]
10%	0.22	0.26	0.21	**0.16** [a]	0.27	**0.27** [b]
20%	0.23	0.28	0.24	**0.17** [a]	0.21	**0.29** [b]
30%	0.25	0.31	0.26	**0.18** [a]	0.29	**0.32** [b]
40%	0.28	0.34	0.29	**0.20**[a]	0.25	**0.34** [b]

*a is the lowest NRMSE

*b is the highest NRMSE

For PFC; as shown in Table 4.2 MissForest has the lowest PFC followed by MICE-CART then MICE, on the other hand side KNN has the highest PFC except for the 40 % missing rate MissRanger has higher PFC.

Table 4.2: Proportion of False Classification (PFC) for the National Health and Nutrition Examination Survey (NHANES) 2007/2008 Dataset

Missing Rate	Methods					
	MICE	MICE-RF	MICE-CART	Miss-Forest	Miss-Ranger	KNN
5%	0.20	0.26	0.18	**0.13** [a]	0.25	**0.40**[b]
10%	0.21	0.27	0.19	**0.14** [a]	0.31	**0.40** [b]
20%	0.25	0.31	0.23	**0.16** [a]	0.36	**0.41** [b]
30%	0.28	0.34	0.26	**0.19** [a]	0.39	**0.41** [b]
40%	0.32	0.38	0.30	**0.22**[a]	**0.46** [b]	0.43

*a is the lowest PFC

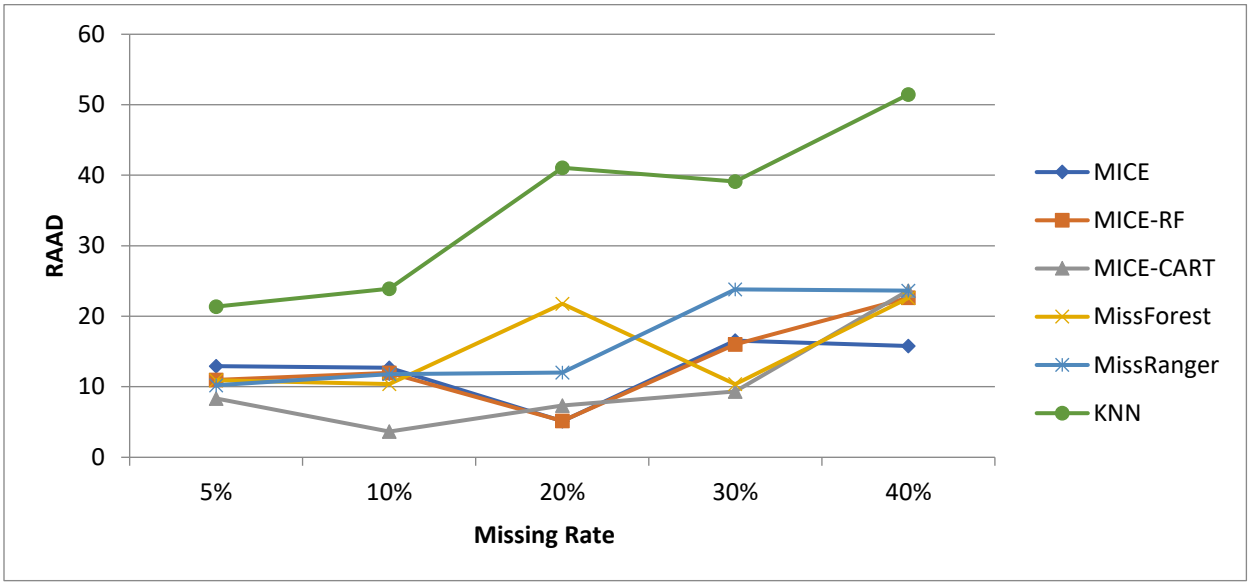
*b is the highest PFC

For evaluating the analyzed model; the analyzed model is assumed to be predicting the weight in the following form

$$\text{weight}=\widehat{B_0} + \widehat{B_1}\ height + \widehat{B_2}\ \text{How do you consider your weight} + \widehat{B_3}\ gender + +\widehat{B_4}\ age + +\widehat{B_5}\ greatest\ weight \quad (4.7)$$

as shown in Figure 4.1; KNN has the highest Relative Average Absolute Difference (RAAD), MICE-CART has the lowest difference in (5%, 10% and 30%) missing rates. The lowest difference at 20% is for MICE-RF and at 40% is for MICE.

Figure 4.1: Relative Average Absolute Difference (RAAD) between estimates of the Model from Complete Data and the Models from Imputed Data for the National Health and Nutrition Examination Survey (NHANES) 2007/2008 Dataset



For the coefficient of determination; as shown in Figure 4.2 MissForest, MissRanger and the complete model have the highest R^2, while KNN has the lowest R^2 followed by MICE-RF

Figure 4.2: Coefficient of determination (R^2) for the National Health and Nutrition Examination Survey (NHANES) 2007/2008 Data

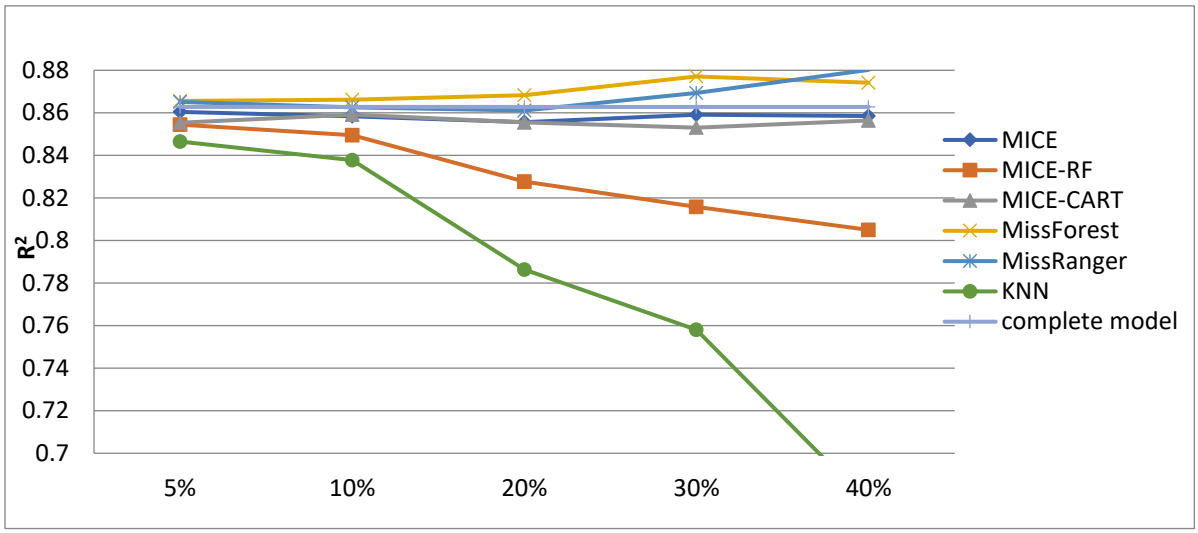
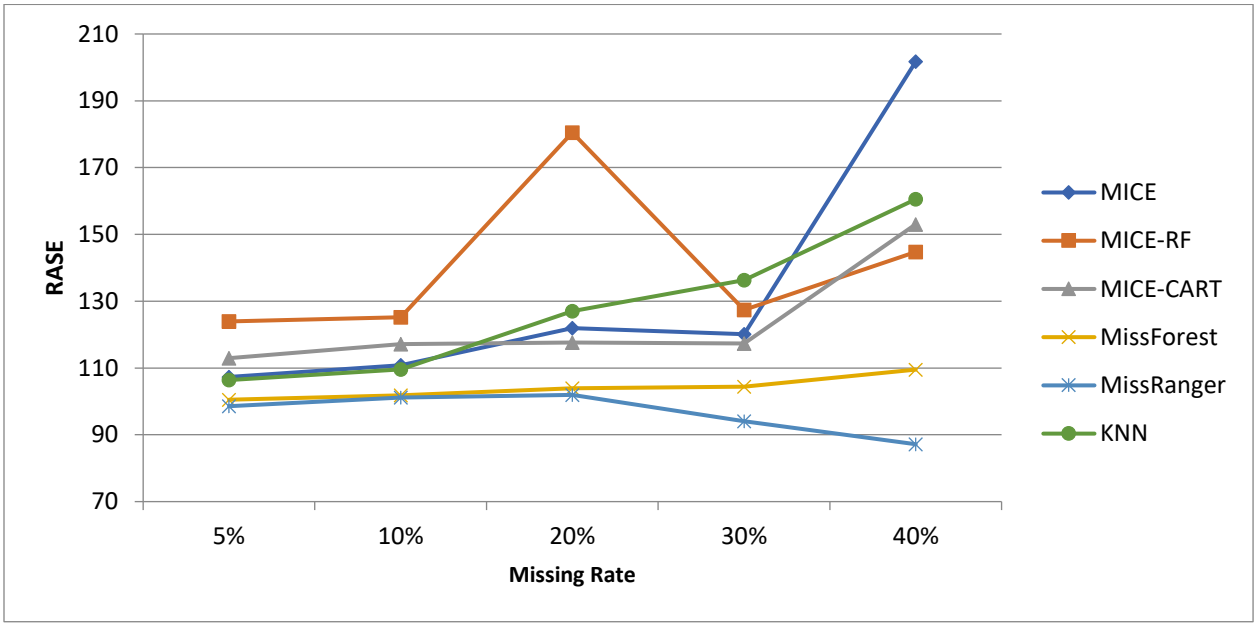


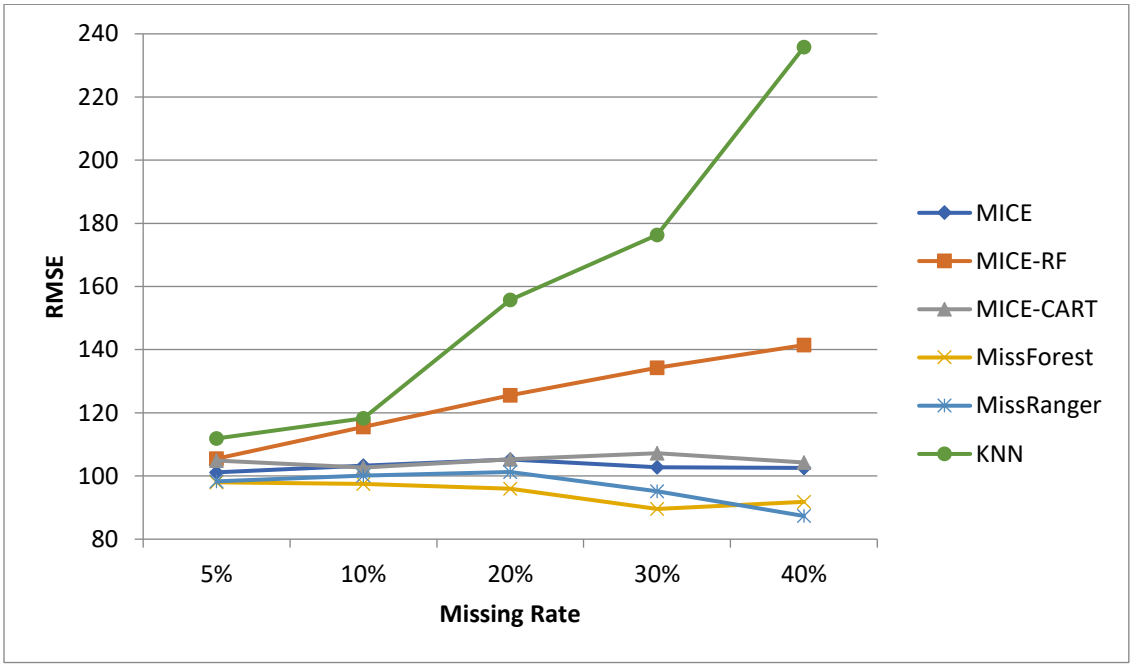
Figure 4.3 shows that MICE-RF has the highest (RASE) among its coefficients till 20% missing rate, while the lowest is for MissRanger and MissForest they are very close to each other.

Figure 4.3: Relative Average Standard Error (RASE) of Estimates of Models from Imputed Data for the National Health and Nutrition Examination Survey (NHANES) 2007/2008 Dataset



For the Relative Mean Square Error (RMSE); Figure 4.4 shows that MissForest and MissRanger has the lowest (RMSE), while KNN has the highest MSE followed by MICE-RF. Table 4.3 provides a summery for the analyzed model results.

Figure 4.4: Relative Mean Square Error (RMSE) of Models from Imputed Data for the National Health and Nutrition Examination Survey (NHANES) 2007/2008 Dataset



Generally, across the analysis model MissForest and MissRanger tend to have the "convenient" results and MICE-RF and KNN tend to have a "inconvenient" result.

Further Investigation is recommended for assessing the performance of the six imputation method under MAR and MNAR assumption, also re-evaluating the performance of MICE-MICE-RF and MICE-CART when increasing the number of the imputed datasets, and MissRanger and MissForest when increasing the number of trees. Beside that it is recommended to re-evaluate the six imputation methods performance when the data contain a categorical variable with large number of categories.

Table 4.3: Summary of Results for the National Health and Nutrition Examination Survey (NHANES) 2007/2008 Dataset

Missing Rate	Methods						
	Criteria	MICE	MICE-RF	MICE-CART	Miss-Forest	Miss-Ranger	KNN
5%	RAAD	12.94	10.97	**8.32** [a]	10.92	10.22	**21.37** [b]
	RASE	107.33	**123.94**	112.95	100.51	**98.56** [a]	106.39
	R^2	86.04%	85.45%	85.54	86.56%	**86.51%**	**84.65%** [a]
	RMSE	101.16	105.49	104.87	**97.99** [a]	98.31	**111.88** [b]
10%	RAAD	12.71	11.97	**3.65** [a]	10.40	11.79	**23.90** [b]
	RASE	110.83	**125.17**	117.12	101.79	**101.11** [a]	109.59
	R^2	85.84%	84.94%	85.92	**86.62%** [a]	86.26%	**83.78%**
	RMSE	103.24	115.50	102.68	**97.54** [a]	100.14	**118.25** [b]
20%	RAAD	**5.12** [a]	5.16	7.36	21.79	12.03	**41.05** [b]
	RASE	121.95	**180.48**	117.62	103.96	**101.96** [a]	126.98
	R^2	85.56%	82.77%	85.55	**86.83%** [b]	86.11%	**78.63%** [a]
	RMSE	105.24	125.60	105.32	**96.01** [a]	101.27	**155.73** [b]
30%	RAAD	16.55	16.03	**9.33** [a]	10.37	23.81	**39.11** [b]
	RASE	120.12	127.45	117.35	104.42	**94.04** [a]	**136.28** [b]
	R^2	85.91%	81.58%	85.30	**87.71%**	86.94%	**75.81%** [a]
	RMSE	102.71	134.32	107.21	**89.57** [a]	95.17	**176.31** [b]
40%	RAAD	**15.79** [a]	22.62	23.66	22.59	23.63	**51.46** [b]
	RASE	**201.71**	144.74	153.03	109.52	**87.18** [a]	160.52
	R^2	85.85%	80.50%	85.64	87.41%	**88.02%**	**67.65%**
	RMSE	102.60	141.44	104.26	91.80	**87.35** [a]	**235.77** [b]

*a is the lowest PFC
*b is the highest PFc

Acknowledgements

The authors would like to thank Dr. Amany Mousa for her kind help and comments on this paper.

References

Amiri, L., Khazaei, M. and Ganjali, M. (2017). A Mixture Latent Variable Model for Modeling Mixed Data in Heterogeneous Populations and Its Applications. *AStA Advances in Statistical Analysis*, 102(1), pp.95-115.

Audigier, V., Husson, F. and Josse, J. (2016). A Principal Component Method to Impute Missing Values for Mixed Data. *Advances in Data Analysis and Classification*, 10(1), pp.5-26.

Ayuyev V.V., Jupin J., Harris P.W., Obradovic Z. (2009) Dynamic Clustering-Based Estimation of Missing Values in Mixed Type Data. In: Pedersen T.B., Mohania M.K., Tjoa A.M. (eds*) Data Warehousing and Knowledge Discovery*. DaWaK 2009. Lecture Notes in Computer Science, 5691. Springer, Berlin, Heidelberg

Bahrami Samani, E., Ganjali, M. and Eftekhari, S. (2010). A Latent Variable Model for Mixed Continuous and Ordinal Responses with Nonignorable Missing Responses: Assessing the Local Influence Via Covariance Structure. *Sankhya B*, 72(1), pp.38-57

Buuren, S. and Groothuis-Oudshoorn, K. (2011). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3).

Buuren, S., Brand, J., Groothuis-Oudshoorn, C. and Rubin, D. (2006). Fully Conditional Specification in Multivariate Imputation. *Journal of Statistical Computation and Simulation*, 76(12), pp.1049-1064.

Dane, S. and R. C., Thool. (2013). Imputation Method for Missing Value Estimation of Mixed-Attribute Data Sets. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(5), pp.729-734

Devi Priya, R. and Sivaraj, R. (2017). Dynamic Genetic Algorithm-Based Feature Selection and Incomplete Value Imputation for Microarray Classification. *Current Science*, 112(1), p.126.

Doove, L., Buurenc, S. and Dusseldorp, E. (2014). Recursive Partitioning for Missing Data Imputation in The Presence of Interaction Effects. *Computational Statistics & Data Analysis*, 72, pp.92-104.

Ganjali, M. (2003). A Model for Mixed Continuous and Discrete Responses with Possibility of Missing Responses. *Journal of Sciences*, Islamic Republic of Iran, 14(1), pp.53-60.

Hörnblad, J. (2013). *Missing Data in the Swedish National Patients Register: Multiple Imputation by Fully Conditional Specification*. Master Thesis. Stockholm University.

Lee, K. and Carlin, J. (2010). Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation. *American Journal of Epidemiology*, 171(5), pp.624-632.

Liao, S., Lin, Y., Kang, D., Chandra, D., Bon, J., Kaminski, N., Sciurba, F. and Tseng, G. (2014). Missing Value Imputation in High-Dimensional Phenomic Data: Imputable or Not, and How? *BMC Bioinformatics*, 15(1).

Little, R. and Schluchter, M. (1985). Maximum Likelihood Estimation for Mixed Continuous and Categorical Data with Missing Values. *Biometrika*, 72(3), p.497.

Raghunathan, T., Lepkowski, J., Van Hoewyk, J. and Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology,* 27(1), pp.85-95.

Rashid Ahmed, M. (2011). *An Investigation of Methods for Missing Data in Hierarchical Models for Discrete Data*. Doctor of Philosophy. University of Waterloo, Canda.

Ren, H. (2012). *Multiple Imputation of High-dimensional Mixed Incomplete Data*. Doctor of Philosophy. University of California.

Sen S., Das M., Chatterjee R. (2018) Estimation of Incomplete Data in Mixed Dataset. In: Sa P., Sahoo M., Murugappan M., Wu Y., Majhi B. (eds) Progress in Intelligent Computing Techniques: Theory, Practice, and Applications. *Advances in Intelligent Systems and Computing*, 518. Springer, Singapore.

Shah, A., Bartlett, J., Carpenter, J., Nicholas, O. and Hemingway, H. (2014). Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. *American Journal of Epidemiology*, 179(6), pp.764-774.

She, X. (2017). *Fractional Imputation for Ordinal and Mixed-type Responses with Missing Observations.* Doctor of Philosophy. University of Waterloo.

Stekhoven, D. and Buhlmann, P. (2011). MissForest-Non-Parametric Missing Value Imputation for Mixed-Type Data. *Bioinformatics,* 28(1), pp.112-118.

Tarsitano, A. and Falcone, M. (2011). Missing-Values Adjustment for Mixed-Type Data. *Journal of Probability and Statistics*, 2011, pp.1-20.

جامعة القاهرة
معهد الدراسات والبحوث الإحصائية

المؤتمر السنوى الثالث والخمسين للإحصاء
وعلوم الحاسب وبحوث العمليات

إحصاء تطبيقى

3-5 ديسمبر 2018

فهرس الإحصاء التطبيقى

تقدير أعداد الفصول والمعلمين بالمدارس الإبتدائية الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦)

سامر محمد سمير محمد سهل*

مستخلص

تعانى مرحلة التعليم الإبتدائى بجمهورية مصر العربية من عدم زيادة الخدمات التعليمية من مبانى تعليمية ومعلمين بنفس معدلات الزيادة فى أعداد الطلبة مما يؤثر على جودة العملية التعليمية ويعيق سبل الإرتقاء بتلك العملية ، ويتضح ذلك فى إرتفاع كثافة الفصول بالمدارس الإبتدائية الحكومية فمثلاً تتعدى كثافة الفصول ٥٠ طالب بالفصل فى كثير من إدارات محافظة القاهرة عام ٢٠١١ ، وتتعدى ٧٠ طالب بالفصل ببعض الإدارات مثل إدارة المرج ٧١ طالب بالفصل وإدارة البساتين ودار السلام ٧٠ طالب بالفصل بينما كثافة الفصل المرجوة ٤٠ طالب بالفصل [11] .

لذا يهتم هذا البحث بتقدير كلا من (أعداد الفصول – أعداد المدارس– أعداد المعلمين) بالمدارس الإبتدائية الحكومية بإدارة المرج بمحافظة القاهرة خلال الفترة من (٢٠١٢-٢٠٢٦) نظراً لأنها أكثر إدارات المحافظة كثافة بالفصول هذا بالإضافة إلى أن أى إسلوب بإدارة المرج يمكن تعميمه على باقى إدارات المحافظة .

ويهدف البحث إلى تقديم نموذج مقترح لتقدير كلا من (أعداد الفصول – أعداد المدارس– أعداد المعلمين) بإدارة المرج ويتضمن تقدير أعداد الطلبة بالمدارس الإبتدائية الحكومية بإدارة المرج خلال فترة التقدير والذى يشمل (تقدير معدل الإستيعاب الصافى – تقدير أعداد السكان عند العمر ٦ سنوات – تقدير أعداد الطلبة بصفوف مرحلة التعليم الإبتدائى) وتقدير (أعداد الفصول – أعداد المدارس– أعداد المعلمين) وتطبيق ذلك النموذج لتقدير الإحتياجات من الخدمات التعليمية بالمدارس الإبتدائية الحكومية بإدارة المرج خلال الفترة من (٢٠١٢– ٢٠٢٦)،

معهد الدراسات والبحوث الإحصائية-جامعة القاهرة

*مدرس مساعد بقسم الرياضة والإحصاء والتأمين بكلية التجارة جامعة أسيوط

وتحديد حجم الفجوة بين المتاح حالياً من خدمات تعليمية والمطلوبة مستقبلياً حتى عام ٢٠٢٦.
وإعتمد بناء النموذج على بيانات مركز الحاسب الألى بوزارة التربية والتعليم خلال الفترة من (١٩٩٢-٢٠١١).

الكلمات المفتاحية : معدل الاستيعاب الصافى ، نماذج التمهيد الإسى ، نموذج تقدير الخدمات التعليمية

(١) مقدمة

تعانى مرحلة التعليم الأساسى بجمهورية مصر العربية وبصفة خاصة المدارس الإبتدائية الحكومية فى العقود الأخيرة من مشاكل تتمثل فى تعدد الفترات الدراسية وتزايد كثافة الفصول وعدم الاستيعاب الكامل للأطفال فى سن القبول بالمرحلة الإبتدائية ، وذلك للزيادة فى أعداد السكان حيث بلغ معدل النمو السكانى بجمهورية مصر العربية بين تعدادى (١٩٩٦-٢٠٠٦) ٢,٠٥% [2] ، وعدم زيادة الخدمات التعليمية من مبانى تعليمية ومعلمين بنفس معدلات الزيادة فى أعداد الطلبة مما يؤثر على جودة العملية التعليمية ويعيق سبل الإرتقاء بتلك العملية [8].

ولا تتضح مشاكل نقص الخدمات التعليمية من مبانى تعليمية ومعلمين عند دراسة الوضع الراهن للاحتياجات من الخدمات التعليمية بالمدارس الإبتدائية على مستوى إجمالى جمهورية مصر العربية حيث أن البيانات المنشورة تفيد أن متوسط كثافة الفصل بالمدارس الإبتدائية بجمهورية مصر العربية (٤٣) طالب بالفصل ومعدل الطلبة إلى المعلمين بالمدارس الإبتدائية الحكومية ٢٧ طالب لكل معلم فى السنة الدراسية (٢٠١٢/٢٠١١) [2] ، وهذا لا يعكس الواقع الفعلى للاحتياجات من الخدمات التعليمية التى ظهرت عند دراسة الوضع الراهن للخدمات التعليمية على سبيل المثال بإدارة المرج التعليمية حيث بلغت على سبيل المثال كثافة الفصل بإدارة المرج ٧١ طالب بالفصل [11] ، وبالتالى فإن التخطيط للإحتياجات من الخدمات التعليمية على المستوى الإجمالي يشوبه الكثير من أوجه القصور، ممايتطلب التخطيط للإحتياجات من الخدمات التعليمية على مستوى إدارات محافظات جمهورية مصر العربية .

حيث أن معظم الدراسات عن التخطيط للإحتياجات من الخدمات التعليمية المستقبلية بجمهورية مصر العربية لم تهتم بالتخطيط للإحتياجات من الخدمات التعليمية على مستوى إدارات محافظات جمهورية مصر العربية .

وفيما يلى يمكن عرض مختصر لأهم الدراسات السابقة التى تناولت التخطيط المستقبلى للاحتياجات من الخدمات التعليمية بجمهورية مصر العربية .

- قامت نعمات تمام ١٩٧٧ باستخدام النماذج الرياضية في تخطيط التعليم بمحافظة الشرقية [7]، وقد توصلت لتقدير الإحتياجات من الخدمات التعليمية (أعداد الفصول- أعداد

المعلمين) بمرحلة التعليم قبل الجامعى (ابتدائي– إعدادي– ثانوي) بمحافظة الشرقية خلال الفترة من (١٩٧٥ – ١٩٨٤).

- قام Diab 1981 بتقديم نموذج رياضى لتقدير الطلبة والمعلمين فى مصر [14] ، وقد توصل لنموذج رياضى تم إستخدامه فى تقدير أعداد الطلبة وأعداد المعلمين بمرحلة التعليم قبل الجامعى (إبتدائى – إعدادى – ثانوى) بمحافظات (القاهرة– الإسكندرية– قنا– البحيرة) خلال الفترة من (١٩٧٦– ١٩٨٦).
- قامت كوثر الحسيني ١٩٨٢ بدراسة المشكلة السكانية والتعليم بالمرحلة الابتدائية في مصر (١٩٦٠ . ٢٠٠٠) [3] وقد توصلت لتقدير (أعداد الطلبة– أعداد الفصول– أعداد المعلمين) مستقبلياً بجمهورية مصر العربية خلال الفترة من (١٩٨٠–٢٠٠٠) .
- قامت ماجدة إبراهيم وآخرون ٢٠٠٢ بتقديم نظرة مستقبلية لمؤشرات الخدمات التعليمية [1] ، وقد توصلت لتقدير أعداد الطلبة والفصول والمدارس والمعلمين بمرحلة التعليم قبل الجامعى (إبتدائى–إعدادى– ثانوى) بجمهورية مصر العربية خلال فترة الدراسة من (٢٠٠١–٢٠٢١).
- قامت تغريد محمود ٢٠٠٤ بالتخطيط لتلبية الاحتياجات الكمية والكيفية لمرحلة التعليم الأساسى بمحافظة القاهرة [12].، توصلت لتقدير الإحتياجات من الخدمات التعليمية من (أعداد الفصول – أعداد المعلمين) بمحافظة القاهرة خلال الفترة من (٢٠٠٣– ٢٠١٧) .

ولم تهتم الدراسات سابقة الذكر بالتخطيط للاحتياجات من الخدمات التعليمية على مستوى إدارات محافظات جمهورية مصر العربية، لذلك فى هذا البحث تم الإهتمام بتقدير الاحتياجات من الخدمات التعليمية (أعداد الفصول– أعداد المدارس – أعداد المعلمين) بالمدارس الإبتدائية الحكومية بإدارة المرج خلال الفترة من (٢٠١٢–٢٠٢٦) من خلال:

أولاً: تقديم نماذج مقترحة لتقدير معدل الاستيعاب الصافى بالمدارس الإبتدائية

بإدارة المرج خلال الفترة من (٢٠١٢–٢٠٢٦) وقد تتطلب ذلك استخدام بيانات أعداد الطلبة بالصف الأول الإبتدائى عند العمر ٦سنوات فى حساب معدل الاستيعاب الصافى بالمدارس الإبتدائية الحكومية خلال الفترة من (١٩٩٢–٢٠١١) كما سيتم إيضاح ذلك لاحقاً .

ثانياً: تقديم نموذج مقترح لتقدير الاحتياجات من الخدمات بالمدارس الإبتدائية الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦) وتطبيق ذلك النموذج لتقدير تلك الاحتياجات من الخدمات التعليمية خلال فترة التقدير وتحديد حجم الفجوة بين المتاح حالياً من خدمات تعليمية والمطلوبة مستقبلياً حتى عام ٢٠٢٦، وتطلب ذلك إستخدام بيانات :

أ- أعداد الطلبة وفقاً للصفوف الدراسية بالمدارس الإبتدائية الحكومية فى السنة الدراسية ٢٠١٢/٢٠١١.

ب- أعداد الطلبة الناجحين والراسبين بين الصفوف الدراسية بالمدارس الإبتدائية الحكومية خلال الفترة من (٢٠٠٨/٢٠٠٧-٢٠١٢/٢٠١١)..

ت- أعداد الفصول والمعلمين بالمدارس الإبتدائية الحكومية فى السنة الدراسية ٢٠١٢/٢٠١١.

حيث تم حساب نسب النجاح والرسوب والتسرب بين الصفوف الدراسية بالمدارس الإبتدائية الحكومية خلال الفترة من (٢٠٠٧-٢٠١١) ، وتحديد حجم الفجوة بين المتاح حالياً من (أعداد الفصول- أعداد المعلمين) فى السنة الدراسية ٢٠١٢/٢٠١١ والمطلوبة مستقبلياً حتى عام ٢٠٢٦ كما سيتم إيضاح ذلك لاحقاً.

وقد تم الحصول على البيانات المطلوبة من مركز الحاسب الألى بوزارة التربية والتعليم .

وفيما يلى عرض للنماذج المقترحة لتقدير معدل الإستيعاب الصافى بالمدارس الإبتدائية الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦) وكذلك النموذج المقترح لتقدير الاحتياجات من الخدمات بالمدارس الإبتدائية الحكومية خلال الفترة من (٢٠١٢-٢٠٢٦) وتطبيق ذلك النموذج لتقدير تلك الاحتياجات من الخدمات التعليمية خلال فترة التقدير

(٢) النماذج المقترحة لتقدير معدل الاستيعاب الصافى بالمدارس الإبتدائية الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦)

و سنعرض النماذج المقترحة لتقدير معدل الاستيعاب الصافى بالمدارس الإبتدائية الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦) وفقاً لمايلى:

يعرف معدل الاستيعاب الصافى بالصف الأول الإبتدائى على أنه نسبة أعداد الطلبة المقيدين بالصف الأول الإبتدائى عند سن القبول الرسمى بمرحلة التعليم الإبتدائى (٦ سنوات)

من إجمالى أعداد السكان عند العمر ٦ سنوات فى السنة t ونشير لذلك المعدل بالرمز $N_6(t)$ ويمكن حسابه وفقا للصيغة التالية [4]:

$$N_6(t) = \frac{E_1(6,t)}{P_6(t)} \qquad (2.1)$$

حيث أن

$E_1(6,t)$ أعداد الطلبة المقيدين بالصف الأول الإبتدائى عند سن ٦ سنوات فى السنة t

$P_6(t)$ أعداد السكان عند العمر ٦سنوات فى السنة t

ووفقا لتعريف معدل الإستيعاب الصافى فى السنة (t) فى المعادلة (1) فإنه تم حساب ذلك المعدل خلال الفترة من (١٩٩٢-٢٠١١) بالمدارس الإبتدائية الحكومية خلال الفترة من (١٩٩٢-٢٠١١) إعتماداً على البيانات التالية [11] :

أ- أعداد الطلبة المقيدين بالصف الأول الإبتدائى بالمدارس الإبتدائية الحكومية عند العمر ٦ سنوات بإدارة المرج خلال الفترة من (١٩٩٢-٢٠١١) .

ب- أعداد السكان عند العمر ٦ سنوات بإدارة المرج خلال الفترة من (١٩٩٢-٢٠١١) ، ونظراً لعدم توافرها فإنه تم تقديرها .

ووفقاً لبيانات معدل الاستيعاب الصافى بالمدارس الإبتدائية الحكومية بإدارة المرج خلال الفترة من (١٩٩٢-٢٠١١) فإن النماذج المقترحة وفيما يلى عرض لتلك النماذج:

١- نموذج التمهيد الأسى Holt linear

يأخذ نموذج التمهيد الأسى (Holt Linear) الشكل الآتى **[6,15,16,20,21]** :

$$S_t = \alpha N_6(t) + (1-\alpha)(S_{t-1} + b_{t-1}) \qquad ,t = 1,2,...,n \qquad (0 \leq \alpha \leq 1) \qquad (2.2)$$

$$b_t = \gamma\left(S_t - S_{t-1}\right) + \left(1-\gamma\right) b_{t-1} \qquad ,t = 1,2,...,n \qquad (0 \leq \gamma \leq 1) \qquad (2.3)$$

$$\hat{N}_6(t') = S_n + b_n t' \qquad , t' \geq 1 \qquad (2.4)$$

حيث إن :

- $N_6(t)$ قيمة معدل الاستيعاب الصافى الفعلية بالمدارس الحكومية فى السنة الدراسية t
- S_t متوسط لقيم معدل الاستيعاب الصافى مرجح بأوزان تتناقص أسياً لجميع القيم السابقة ، ويشمل إتجاه السلسلة فى الماضى
- b_t قيمة تمهيد الاتجاه الخطي عند الزمن t
- S_{t-1} متوسط معدل الاستيعاب الصافى مرجح بأوزان تتناقص أسيا لجميع القيم السابقة بدون معدل الاستيعاب الصافى عند الزمن t ويشمل إتجاه السلسلة فى الماضى
- b_{t-1} قيمة تمهيد الإتجاه الخطي عند الزمن $t-1$
- $\hat{N}_6(t')$ t' قيمة معدل الاستيعاب الصافى بالمدارس الحكومية المتنبأ بها خلال الفترة المستقبلية
- فترة السلسلة الزمنية (الفترة من ١٩٩٢-٢٠١١)

ومعلمات النموذج هما (α) : معامل التمهيد ، وكذلك (γ) معامل الإتجاه ، ويمكن تقدير المعلمتين باستخدام طريقة التجربة والخطأ (trials and error) [20] ،ويتم إختيار معلمات النموذج بإستخدام البرنامج الإحصائى SPSS فيما يسمى ببحث الشبكة (Grid Search) **. [20]**

٢- نموذج التمهيد الأسى المضاعف Brown

يأخذ نموذج التمهيد الأسى المضاعف (Brown) الشكل الآتي :[14,16,19,21]

$$S_t^{(1)} = \alpha N_6(t) + (1-\alpha) S_{t-1}^{(1)} \qquad , t = 1,2,\ldots,n \qquad (0 \leq \alpha \leq 1) \qquad (2.5)$$

$$S_t^{(2)} = \alpha S_t^{(1)} + (1-\alpha) S_{t-1}^{(2)} \qquad , t = 1,2,\ldots,n \qquad (0 \leq \alpha \leq 1) \qquad (2.6)$$

$$\hat{N}_6(t') = (2 + \frac{\alpha t'}{1-\alpha})S_n^{(1)} - (1 + \frac{\alpha t'}{1-\alpha})S_n^{(2)} \qquad (2.7)$$

حيث إن :

- تمهيد أسى بسيط (من الدرجة الاولى) لمعدل الاستيعاب الصافى $S_t^{(1)}$

تمهيد أسى مضاعف (من الدرجة الثانية) لمعدل الاستيعاب الصافى $S_t^{(2)}$

ومعلمة النموذج هى (α) ويمكن تقدير المعلمة باستخدام طريقة التجربة والخطأ (trials and error) [19] ، SPSS فيما ببحث الشبكة (Grid Search) **[20]**.

نموذج التمهيد الأسى Damped linear

يأخذ نموذج التمهيد الأسى (Damped linear) الشكل الآتي **[20 ,16]**:

$$S_t = \alpha N_6(t) + (1-\alpha)(S_{t-1} + \phi b_{t-1}) \qquad ,t = 1,2,..,n \quad (0 \le \phi, \alpha \le 1) \qquad (2.8)$$

$$b_t = \gamma (S_t - S_{t-1}) + (1-\gamma)\phi b_{t-1} \qquad ,t = 1,2,..,n \quad (0 \le \gamma \le 1) \qquad (2.9)$$

$$\hat{N}_6(t') = S_n + b_n \sum_{i=1}^{t'} \phi^i \qquad t' \geqslant 1 \qquad (2.10)$$

ومعلمات النموذج هم (α) : معامل التمهيد ، وكذلك (γ) معامل الإتجاه ،السابق إيضاحهما بنموذج التمهيد الأسى (Holt Linear) ، بالإضافة إلى معامل تخفيض أثر اتجاه السلسلة ϕ ، ويمكن تقدير المعلمات باستخدام طريقة التجربة والخطأ (trials and error) ، ويتم إختيار معلمات النموذج التى بإستخدام البرنامج الإحصائىSPSS فيما يسمى ببحث الشبكة (Grid Search).

وقد وجد أن نموذج التمهيد الأسى المضاعف Brown هو الأفضل لتقدير معدل الإستيعاب الصافى للذكور بإدارة المرج بينما نموذج التمهيد الأسى Damped linear هو الأفضل لتقدير معدل الإستيعاب الصافى للإناث حيث لهما أقل حجم أخطاء وفقاً لمقياس متوسط مربعات الأخطاء (MSE) وموضح ذلك بجدول (١-١) بالملحق .

(٣) النموذج المقترح لتقدير الإحتياجات من الخدمات بالمدارس الإبتدائية الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦)

و سنعرض النموذج المقترح لتقدير الإحتياجات من الخدمات التعليمية بالمدارس الإبتدائية الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦) وهو نموذج كوريا للتخطيط التعليمى (Correa,1969) **[18]** والذى يتطلب:

١- مدخلات النموذج :

أ- تقدير معدل الاستيعاب الصافى بالمدارس الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦).

ب- تقدير أعداد السكان عند العمر ٦ سنوات بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦).

ت- تقدير نسب النجاح والرسوب والتسرب بين الصفوف الدراسية بالمدارس الإبتدائية الحكومية خلال الفترة من (٢٠١٢-٢٠٢٦).

ث- أعداد الطلبة بصفوف مرحلة التعليم الإبتدائى بالمدارس الحكومية بإدارة المرج فى السنة الدراسية ٢٠١٢/٢٠١١.

٢- مخرجات النموذج :

أ- تقدير أعداد الطلبة بصفوف مرحلة التعليم الإبتدائى بالمدارس الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦).

ب- تقدير أعداد الفصول المطلوبة بالمدارس الإبتدائية الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦).

ت- تقدير أعداد المعلمين المطلوبين بالمدارس الإبتدائية الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦).

وقد إعتمد البحث فى تطبيق ذلك النموذج على الفروض الاتية:

- ثبات نسب النجاح ونسب الرسوب ونسب التسرب بين الصفوف الدراسية بمرحلة التعليم الإبتدائى بالمدارس الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦)، وذلك بإعتبارها متوسط لنسب النجاح ونسب الرسوب ونسب التسرب خلال الفترة من (٢٠١٢/٢٠١١-٢٠٠٨/٢٠٠٧).

- المستجدين على النظام التعليمى بمرحلة التعليم الإبتدائى هم فقط المستجدين بالصف الأول الإبتدائى.
- لايوجد عودة للطلبة بعد تسربهم من أى صف دراسى.
- كثافة الفصل المرجوة ٤٠ طالب بالفصل.

وقد أوضحنا سابقا كيفية تقدير أحد مدخلات النموذج وهو تقدير معدل الإستيعاب الصافى بالمدارس الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦) وسوف نقدم كيفية تقدير وحساب باقى مدخلات النموذج ثم نقوم بإيضاح كيفية إستخدام تلك المدخلات فى تقدير مخرجات النموذج كما يلى:

أولاً: كيفية تقدير وحساب مدخلات النموذج

أ- تقدير أعداد السكان عند العمر ٦ سنوات بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦)

تم تقدير اعداد السكان عند العمر ٦ سنوات بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦) إعتمادا على الاسقاطات السكانية لمحافظات جمهورية مصر العربية خلال الفترة من (٢٠٠٦-٢٠٣١) التى قدمتها (هند عطية) عام ٢٠٠٩ [10] باستخدام طريقة مكونات الأفواج Component (The cohort Method) وفقاً لأربعة بدائل أو سيناريوهات متعلقة بمستويات الوفيات والإنجاب :

وفى البحث تم إختيار أفضل بديل من تلك البدائل بمقارنة تقديرات سكان محافظة القاهرة ٢٠١١ بتلك الدراسة مع بيانات الكتاب الإحصائى السنوى الصادر من الجهاز المركزى للتعبئة العامة والإحصاء عام ٢٠١١[٢] ، وذلك بإستخدام نسبة الخطأ المطلقِ (Percentage Error Absolute) وفقا للمعادلة الأتية :

$$\text{نسبة الخطأ المطلق} = \frac{|\text{إسقاطات أعداد السكان} - \text{عدد السكان فى كتاب الإحصاء السنوى } \mathbf{2011}|}{\text{عدد السكان فى كتاب الإحصاء السنوى } \mathbf{2011}} * \mathbf{100} \qquad (3.1)$$

وقد وجد أن البديل الثالث (انخفاض الوفيات وثبات الانجاب) هو أنسب بديل سكانى، وقد تم إستخدام إسقاطات أعداد السكان بالفئات العمرية الخمسية (٠-٤) ، (٥-٩) ، (١٠-١٤) ، (١٥-١٩) ، (٢٠-٢٤) لتلك الدراسة المتعلقة بمحافظة القاهرة وفقا للبديل الثالث

خلال فترة التقدير من (٢٠١٢ -٢٠٢٦) فى تقدير أعداد السكان عند العمر ٦سنوات بإدارة المرج والموضحة وفقا لطريقة النسبة (Ratio Method) .

ب- تقدير نسب النجاح والرسوب والتسرب بين الصفوف الدراسية بالمدارس الإبتدائية الحكومية خلال الفترة من (٢٠١٢-٢٠٢٦) ، وقد أوضحنا ذلك عند عرض فروض النموذج.

ت- أعداد الطلبة بصفوف مرحلة التعليم الإبتدائى بالمدارس الحكومية بإدارة المرج فى السنة الدراسية ٢٠١٢/٢٠١١

تم الحصول على بيانات أعداد الطلبة بصفوف مرحلة التعليم الإبتدائى بالمدارس الحكومية بإدارة المرج فى السنة الدراسية ٢٠١٢/٢٠١١ من مركز الحاسب الألى بوزارة التربية والتعليم.

ثانياً: كيفية تقدير مخرجات النموذج

أ- تقدير أعداد الطلبة بصفوف مرحلة التعليم الإبتدائى بالمدارس الحكومية بإداراة المرج خلال الفترة من (٢٠١٢-٢٠٢٦)

تم تقدير أعداد الطلبة بالمدارس الإبتدائية الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦) وفقاً لنموذج كوريا إعتماداً على المصفوفات الآتية :

١- متجه أعداد الطلبة $E'(t)$ الذى يشمل أعداد الطلبة بصفوف المرحلة الابتدائية بالإضافة إلى تقدير أعداد السكان عند العمر ٦ سنوات.

٢- مصفوفة التدفقات الطلابية Z'_{ij} التى تشمل نسب النجاح والرسوب والتسرب بين الصفوف الدراسية بالمرحلة الإبتدائية بالإضافة إلى معدل الاستيعاب الصافى المقدر $\hat{N}_6(t')$ والموضحه بجدول (١-٢) بالملحق وتأخذ المصفوفة الشكل التالى:

$$Z'_{ij} = \begin{bmatrix} \hat{N}_6(t) & 0 & 0 & 0 & 0 & 0 & 0 \\ r_{11} & s_{12} & 0 & 0 & 0 & 0 & w_1 \\ 0 & r_{22} & s_{23} & 0 & 0 & 0 & w_2 \\ 0 & 0 & r_{33} & s_{34} & 0 & 0 & w_3 \\ 0 & 0 & 0 & r_{44} & s_{45} & 0 & w_4 \\ 0 & 0 & 0 & 0 & r_{55} & s_{56} & w_5 \\ 0 & 0 & 0 & 0 & 0 & r_{66} & w_6 \end{bmatrix} \qquad (3.2)$$

و تم تقدير أعداد الطلبة بمرحلة التعليم الإبتدائى وفقا للمعادلة الأتية :

$$E(t+1)=E'(t)*Z'_{ij}=\left[E_1(t+1)\ E_2(t+1)\ E_3(t+1)\ E_4(t+1)\ E_5(t+1)\ E_6(t+1)\ W(t)\right] \quad (3.3)$$

حيث أن

أعداد الطلبة بالصف الدراسى i فى السنة t	$E_i(t)$
أعداد الطلبة بالصف الدراسى j فى السنة $t+1$	$E_j(t+1)$
أعداد الطلبة بالصف الدراسى j فى السنة t	$E_j(t)$
نسبة النجاح بين الصف i والصف j فى السنة t	$s_{ij}(t)$
نسبة الرسوب بالصف الدراسى j فى السنة t	$r_{jj}(t)$

ب- تقدير أعداد الفصول المطلوبة بالمدارس الإبتدائية الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦)

تم تقدير اعداد الفصول المطلوبين مستقبلياً بالمدارس الإبتدائية الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦) إعتمادا على تقديرات اعداد الطلبة الموضحة بالخطوات السابقة و كثافة الفصول المرجوة مستقبليا وذلك وفقا للمعادلة الأتية [4]:

$$CN(t)=\frac{\sum_{i=1}^{6}E_i(t)}{CS} \quad (3.4)$$

حيث إن:

تقديرات أعداد الفصول المطلوبة مستقبلياً فى السنة الدراسية t	$CN(t)$
كثافة الفصل	CS

وقد تم إفتراض أن كثافة الفصل المأمول بجمهورية مصر العربية فى هذا البحث ٤٠ طالب ، حيث نص القرار الوزارى رقم (١٤٨) لسنة ٢٠٠٠ على ضرورة تلافى التباين فى الفصول بالمدارس الحكومية وذلك بالوصول بكثافة الفصل إلى ٤٠ طالب بالتعليم الأساسى.

وتم تحديد حجم الفجوة بين أعداد الفصول المطلوبة خلال الفترة من (٢٠١٢-٢٠٢٦) وأعداد الفصول المتاحة حالياً فى السنة الدراسية ٢٠١٢/٢٠١١ بالمدارس الإبتدائية الحكومية بإدارات محافظة القاهرة وفقاً للعلاقة الآتية:

$$(Gap\ CN(t')) = CN(t') - CN(2011) \qquad (t' = 2012,2013,...,2026) \qquad (3.5)$$

حيث إن :

- الفجوة فى أعداد الفصول المطلوبة مستقبلياً والمتاحة حالياً
  $Gap\ CN(t')$ أعداد الفصول المطلوبة بالسنة الدراسية المستقبلية t'
  $CN(t')$
- أعداد الفصـــول المتاحة فى الوضـــع الراهن فى الســـنة الدراســـية ٢٠١١
  $CN(2011)$

ت- تقدير أعداد المعلمين المطلوبين بالمدارس الإبتدائية الحكومية بإدارات محافظة القاهرة خلال الفترة من (٢٠١٢-٢٠٢٦)

تم تقدير أعداد المعلمين المطلوبين مستقبلياًبالمدارس الإبتدائية الحكومية بالإدارات التعليمية بمحافظة القاهرة خلال الفترة من (٢٠١٢-٢٠٢٦) اعتماداً على التقديرات المتعلقة بأعداد الفصول المطلوبة بالمرحلة الابتدائية وعدد ساعات التدريس لكل مادة ، وكذلك أعداد الساعات المخصصة لكل معلم حيث تم تقدير أعداد المعلمين المطلوبين مستقبلياً لكل مادة دراسية k وفقاً للمعادلة الأتية:

$$Teacher(t,k) = \frac{CN(t) * H(k)}{L(k)} \qquad (k = 1,2,...,7) \qquad (3.6)$$

حيث إن :

- k تشير للمادة الدراسية (لغة عربية- رياضيات-اللغة الإنجليزية- العلوم- الدراسات الإجتماعية- النشاط الفنى- النشاط الرياضى)
- أعداد المعلمين المطلوبين للمادة الدراسية k فى السنة الدراسية t
  $Teacher(t,k)$
- متوسط عدد ساعات التدريس لكل فصل فى الوحدة الزمنية $H(k)$
  (يوم . أسبوع....) للمادة الدراسية k
- متوسط عدد ساعات التدريس المخصصة للمعلم في الوحدة الزمنية $L(k)$
  (يوم . أسبوع....) للمادة الدراسية k

وقد تم الحصول على عدد ساعات الدراسة الإسبوعية بمرحلة التعليم الإبتدائى بالصفوف الدراسية من الخطة الدراسية للعام الدراسى (٢٠١١/٢٠١٢)

كما حددت وزارة التربية والتعليم النصاب القانونى للحصص لمعلمى المرحلة الابتدائية ٢٤ حصة اسبوعياً .

وتم تجميع أعداد المعلمين المطلوبين بكل مادة والمحسوبة وفقاً للمعادلة (3.6) ، وذلك للحصول على أعداد المعلمين المطلوبين بمرحلة التعليم الإبتدائى.

وتم تحديد حجم الفجوة بين أعداد المعلمين المطلوبين مستقبلياًخلال الفترة من (٢٠١٢-٢٠٢٦) ، وأعداد المعلمين الموجودين حالياً فى السنة الدراسية ٢٠١١/٢٠١٢ بالمدارس الإبتدائية الحكومية بإدارات محافظة القاهرة وفقاً للمعادلة الآتية:

$$\text{(Gap Teacher(t'))} = \text{Teacher(t')} - \text{Teacher(2011)} \quad (t' = 2012, 2013, \ldots, 2026) \qquad (3.7)$$

حيث إن:

- (Gap Teacher(t')) الفجوة فى أعداد المعلمين المطلوبين مستقبلياً والمتاحة حالياً
- Teacher(t') أعداد المعلمين المطلوبين بالسنة الدراسية المستقبلية m
- Teacher(2011) أعداد المعلمين المتاحين فى السنة الدراسية ٢٠١١

وقد تم تطبيق النموذج الموضح سابقاً لتقدير الإحتياجات من الخدمات التعليمية بالمدارس الإبتدائية الحكومية بإدارة المرج ، وتم تقدير أعداد المدارس المطلوب بناءها مستقبلياً بالمدارس الإبتدائية الحكومية بإدارة المرج إعتماداً على حجم الفجوة بين أعداد الفصول المطلوبة خلال الفترة من (٢٠١٢-٢٠٢٦) وأعداد الفصول المتاحة حالياً فى سنة الاساس (٢٠١١) بالمدارس الإبتدائية الحكومية بإداراة المرج ، ووفقا لمعايير الهيئة العامة للتخطيط العمرانى [5]التى أوصت أن تتضمن كل مدرسة إبتدائية ٢٤ فصل كحد مرغوب فيه ، وموضح بالجداول من (١-٣) إلى (١-٥) بالملحق تقدير أعداد الطلبة وأعداد الفصول وأعداد المدارس وأعداد المعلمين خلال الفترة من (٢٠١٢-٢٠٢٦) على بإدارة المرج.

(٤) نتائج البحث

١. تقدير معدل الاستيعاب الصافى بالمدارس الإبتدائية الحكومية بإدارة المرج
٢. تقدير أعداد السكان عند العمر ٦ سنوات الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦) .
٣. تقدير نسب النجاح والرسوب والتسرب بين الصفوف الدراسية بالمدارس الإبتدائية الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦).
٤. تطبيق نموذج كوريا للتخطيط التعليمى لتقدير :

أ-أعداد الطلبة بصفوف مرحلة التعليم الإبتدائى بالمدارس الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦).

ب- أعداد الفصول المطلوبة بالمدارس الإبتدائية الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦).

ت- أعداد المعلمين المطلوبين بالمدارس الإبتدائية الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦).

٥. تحديد حجم الفجوة بين أعداد الفصول أعداد المعلمين المطلوبين بالمدارس الإبتدائية الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦) وأعداد تلك الخدمات المتاحة حالياً.

٦. تقدير أعداد المدارس المطلوب بناءها مستقبلياً بالمدارس الإبتدائية الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦) .

المراجع

أولا : المراجع العربية

1. إبراهيم ، ماجدة والقصاص، عبد الحميد والريس، أماني (٢٠٠٢) :" نظرة مستقبلية لمؤشرات الخدمات التعليمية". **سلسلة قضايا التخطيط والتنمية.** رقم ١٥٥ . معهد التخطيط القومي بجمهورية مصر العربية.

2. الجهاز المركزى للتعبئة العامة والإحصاء .(٢٠١٣). "**كتاب الإحصاء السنوى**". القاهرة جمهورية مصر العربية.

3. الحسينى، كوثر محمد (١٩٨٢) ." **المشكلة السكانية والتعليم بالمرحلة الابتدائية في مصر (١٩٦٠ - ٢٠٠٠)** " . رسالة ماجستير . معهد الدراسات والبحوث الإحصائية، جامعة القاهرة، جمهورية مصر العربية.

4. الدليل الفني. (٢٠٠٩). "**المؤشرات القومية للتعليم في مصر** " . وزارة التربية والتعليم . جمهورية مصر العربية.

5 . الهيئة العامة للتخطيط العمرانى .(٢٠١٤) : "**دليل المعدلات والمعايير التخطيطية للخدمات** مصر العربية.

6. برى ، عدنان ماجد عبدالرحمن .(٢٠٠٢): "**طرق التنبؤ الإحصائى ، الجزء الأول**".جامعة الملك سعود.المملكة العربية السعودية.

7. تمام، نعمات مرسي، (١٩٧٧) :" **استخدام النماذج الرياضية في تخطيط التعليم محافظة الشرقية** ". رسالة ماجستير، كلية الاقتصاد والعلوم السياسية ، قسم الإحصاء .جامعة القاهرة.جمهورية مصر العربية.

8. زكرى،لورنس بسطا.(٢٠٠٨)."**كثافة الفصول في التعليم الأساسي (المشكلة وأساليب مواجهتها)**".المركز القومي للبحوث التربوية.جمهورية مصر العربية.

9. شعراوى، سمير مصطفى.(٢٠٠٥). "**مقدمة فى التحليل الحديث للسلاسل الزمنية**" .الملك عبدالعزيز .مركز النشر العلمى. الطبعة الأولى.

10. عطية، هند عطية سيد.(٢٠٠٩)." **إستخدام الأسلوب التجميعى لإعداد الإسقاطات السكانية بجمهورية مصر العربية**" .رسالة ماجيستير .كلية التجارة وإدارة الأعمال. قسم الإحصاء ورياضيات التأمين .جامعة حلوان. جمهورية مصر العربية.

11. محمد سمير ، سامر (٢٠١٦) "**تقدير أعداد طلاب مرحلة التعليم الأساسى لتقدير الإحتياجات من الخدمات التعليمية فى جمهورية مصر العربية خلال الفترة من ٢٠١١-٢٠٢٦** ". رسالة ماجستير غير منشورة . كلية التجارة وإدارة الأعمال . قسم الإحصاء ورياضيات التأمين. جامعة حلوان. جمهورية مصر العربية.

12. محمود، تغريد محمد (٢٠٠٤) ." **التخطيط لتلبية الاحتياجات الكمية والكيفية لمرحلة التعليم الأساسي بمحافظة القاهرة**".رسالة ماجستير، كلية التربية، قسم أصول التربية، جامعة حلوان، جمهورية مصر العربية.

ثانيا : المراجع الأجنبية

13. Correa, H .(1969). "**Quantitative Methods of Educational Planning** ". International Text Book Company.Penncylavania.USA

14. Diab, I.M. (1981)."**The Development of Mathematical Planning Model for The Estimation of School Enrollment and Teaching Staff Demand in Egypt**".Ph.D.The Pennsylvania State University.USA

15. Everette S. Gardner Jr. (2006). "Exponential smoothing: The state of the art—Part II"**. International Journal of Forecasting**. Vol.22, PP 637-666.

16. Gardner, D. E . (1981)."Weight Factor Selection in Double Exponential Smoothing Enrollment Forecasts". **Research in Higher Education** .Vol .14. N.1, PP 49-56.

17. Gaynor, P.E ,Kirkpatrick, R.C .(1994)."**Introduction to Time- Series Modeling and Forecasting in Business and Economics** ".McGraw-Hill, Inc.USA.

18. Mimmack,G.M.,Mayer,D.H.(2001**). "Introductory Statistics for Business".** Pearson Education South Africa. South Africa.

19. Montgomery,D.C.,LynwoodA.J,Gardiner,J.S. .(1990)." **Forecasting and Time Series Analysis**". second edition. McGraw- Hill, Inc.USA.

20. Yaffee,R,McGee,M.(2011):"Analysis and Forecasting with Application SAS&SPSS " Academic press, Inc . New York .USA.

21. Zuhaimy,I .(2011):"Genetic Algorithm Parameter in Double Exponential smothing " **Australian Journal of Basic and Applied Science**.Vol.5. N.7, PP (1174-1180).

ملحق (١)

فى هذا الملحق سوف نقدم :

١- قيمة MSE المحسوبة للنماذج المقترحة لتقدير معدل الاستيعاب الصافى للذكور والإناث .

٢- تقدير معدل الإستيعاب الصافى بالمدارس الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦).

٣- تقدير أعداد الفصول وأعداد المدارس المطلوب بناءها و أعداد المعلمين المطلوبين بالمدارس الحكومية الإبتدائية بإدارات محافظة القاهرة خلال الفترة من (٢٠١٢-٢٠٢٦) .

وفيما يلى سنتناول ذلك بالتفصيل :

١- قيمة MSE المحسوبة للنماذج المقترحة لتقدير معدل الاستيعاب الصافى للذكور والإناث .

يوضح الجدول التالى قيمة MSE المحسوبة للنماذج المقترحة لتقدير معدل الاستيعاب الصافى للذكور والإناث

جدول (١-١) قيمة MSE المحسوبة للنماذج المقترحة لتقدير معدل الاستيعاب الصافى للذكور والإناث

النموذج	MSE	
	الذكور	الإناث
التمهيد الأسى (Holt)	.00241	.00298
التمهيد الأسى (Brown)	.00235	.00347
التمهيد الأسى (Damped)	.00255	.00295

٢- تقدير معدل الإستيعاب الصافى بالمدارس الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦).

يوضح الجدول التالى تقدير معدل الإستيعاب الصافى بالمدارس الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦) :

جدول (١-٢): تقدير معدل الإستيعاب الصافى بالمدارس الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦)

السنة	تقديرات معدل الإستيعاب الصافى بالمدارس الحكومية	
	ذكور	إناث
2012	0.600	0.596
2013	0.603	0.605
2014	0.607	0.611
2015	0.610	0.615
2016	0.614	0.618
2017	0.618	0.620
2018	0.621	0.621
2019	0.625	0.622
2020	0.629	0.623
2021	0.632	0.623

معهد الدراسات والبحوث الإحصائية-جامعة القاهرة

2022	0.636	0.624
2023	0.640	0.624
2024	0.643	0.624
2025	0.647	0.624
2026	0.650	0.624

٣- تقدير أعداد الفصول وأعداد المدارس المطلوب بناءها و أعداد المعلمين المطلوبين بالمدارس الحكومية الإبتدائية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦)

يوضح الجداول التالية تقدير أعداد الطلبة وأعداد الفصول وأعداد المدارس وأعداد المعلمين بالمدارس الإبتدائية خلال الفترة من (٢٠١٢-٢٠٢٦) بإدارة المرج على:

جدول (١-١): تقدير أعداد الطلبة (ذكور- اناث) بالمدارس الإبتدائية الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦)

2012							
	الصف الأول	الصف الثانى	الصف الثالث	الصف الرابع	الصف الخامس	الصف السادس	الاجمالى
الذكور	5294	5500	4442	4029	3833	4140	27238
الاناث	5227	4924	4125	4002	3580	3750	25608
الاجمالى	10521	10424	8567	8031	7413	7890	52846
2013							
	الصف الأول	الصف الثانى	الصف الثالث	الصف الرابع	الصف الخامس	الصف السادس	الاجمالى
الذكور	5419	5289	5282	4296	3835	4009	28130
الاناث	5394	5180	4713	4021	3751	3631	26691
الاجمالى	10813	10469	9996	8317	7586	7640	54821
2014							
	الصف الأول	الصف الثانى	الصف الثالث	الصف الرابع	الصف الخامس	الصف السادس	الاجمالى
الذكور	5541	5405	5115	5083	4077	3991	29210
الاناث	5538	5348	4968	4577	3774	3774	27980
الاجمالى	11079	10753	10083	9659	7851	7765	57190
2015							
	الصف الأول	الصف الثانى	الصف الثالث	الصف الرابع	الصف الخامس	الصف السادس	الاجمالى
الذكور	5664	5526	5217	4971	4801	4201	30380
الاناث	5666	5492	5131	4835	4282	3809	29214
الاجمالى	11330	11018	10348	9806	9083	8010	59594
2016							
	الصف الأول	الصف الثانى	الصف الثالث	الصف الرابع	الصف الخامس	الصف السادس	الاجمالى

الذكور	5788	5649	5334	5059	4734	4871	31435
الاناث	5782	5619	5269	4997	4531	4275	30473
الاجمالى	11570	11268	10603	10055	9266	9146	61908
	2017						
	الصف الأول	الصف الثانى	الصف الثالث	الصف الرابع	الصف الخامس	الصف السادس	الاجمالى
الذكور	5910	5773	5453	5171	4811	4915	32032
الاناث	5884	5734	5391	5133	4686	4548	31376
الاجمالى	11794	11507	10844	10303	9497	9463	63408
	2018						
	الصف الأول	الصف الثانى	الصف الثالث	الصف الرابع	الصف الخامس	الصف السادس	الاجمالى
الذكور	6023	5895	5572	5286	4916	4989	32680
الاناث	5980	5835	5503	5252	4814	4716	32101
الاجمالى	12003	11730	11075	10538	9730	9705	64781
	2019						
	الصف الأول	الصف الثانى	الصف الثالث	الصف الرابع	الصف الخامس	الصف السادس	الاجمالى
الذكور	6142	6008	5690	5402	5025	5093	33359
الاناث	6073	5931	5600	5361	4927	4850	32742
الاجمالى	12215	11939	11290	10763	9952	9942	66101
	2020						
	الصف الأول	الصف الثانى	الصف الثالث	الصف الرابع	الصف الخامس	الصف السادس	الاجمالى
الذكور	6262	6126	5799	5516	5135	5205	34043
الاناث	6163	6023	5692	5457	5030	4965	33330
الاجمالى	12424	12149	11491	10973	10165	10170	67373
	2021						
	الصف الأول	الصف الثانى	الصف الثالث	الصف الرابع	الصف الخامس	الصف السادس	الاجمالى
الذكور	6382	6246	5914	5623	5245	5319	34728
الاناث	6163	6112	5780	5547	5120	5070	33792
الاجمالى	12546	12358	11694	11169	10364	10389	68520
	2022						
	الصف الأول	الصف الثانى	الصف الثالث	الصف الرابع	الصف الخامس	الصف السادس	الاجمالى
الذكور	6492	6366	6029	5733	5346	5433	35400
الاناث	6327	6114	5866	5633	5204	5163	34307
الاجمالى	12820	12480	11895	11366	10550	10595	69707
	2023						
	الصف الأول	الصف	الصف الثالث	الصف الرابع	الصف الخامس	الصف السادس	الاجمالى
الذكور	6603	6476	6146	5845	5451	5540	36061
الاناث	6404	6274	5870	5716	5285	5249	34798
الاجمالى	13006	12750	12015	11562	10737	10788	70859

تابع جدول (١-٣)

	2024					
	الصف الأول	الصف	الصف الثالث	الصف الرابع	الصف الخامس	الصف السادس	الاجمالى
الذكور	6714	6586	6252	5958	5558	5649	36717
الاناث	6479	6351	6020	5723	5364	5331	35268
الاجمالى	13193	12938	12272	11681	10922	10980	71985
	2025						
	الصف الأول	الصف	الصف الثالث	الصف الرابع	الصف الخامس	الصف السادس	الاجمالى
الذكور	6826	6697	6358	6062	5665	5759	37369
الاناث	6554	6426	6096	5865	5372	5411	35723
الاجمالى	13380	13124	12455	11927	11037	11170	73092
	2026						
	الصف الأول	الصف	الصف الثالث	الصف الرابع	الصف الخامس	الصف السادس	الاجمالى
الذكور	6939	6809	6466	6165	5765	5871	38015
الاناث	6629	6501	6168	5941	5501	5426	36165
الاجمالى	13568	13310	12634	12106	11266	11297	74180

جدول (١-٢) : تقدير أعداد الفصول والمدارس المطلوب بناءها بالمدارس الإبتدائية الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦)

السنة	أعداد الفصول المطلوبة كثافة الفصل ٤٠	الفجوة فى أعداد الفصول	الزيادة السنوية فى أعداد الفصول	أعداد المدارس المطلوب بناءها مستقبلا	الزيادة السنوية فى أعداد المدارس المطلوب بناءها مستقبلا
2012	1321	602		25	
2013	1371	651	50	27	2
2014	1430	711	60	30	3
2015	1490	771	60	32	3
2016	1548	829	58	35	2
2017	1585	866	37	36	2
2018	1620	901	35	38	1
2019	1652	933	33	39	1
2020	1684	965	32	40	1
2021	1713	994	29	41	1
2022	1743	1024	30	43	1
2023	1771	1052	28	44	1
2024	1800	1081	29	45	1
2025	1827	1108	27	46	1
2026	1855	1136	28	47	1
	أعداد الفصول الفعلية ٢٠١١	719			

جدول (١-٣) : تقدير أعداد المعلمين بالمدارس الإبتدائية الحكومية بإدارة المرج خلال الفترة من (٢٠١٢-٢٠٢٦)

السنة	اللغة العربية والتربية الدينية	الرياضيات	اللغة الإنجليزية	العلوم	الدراسات الإجتماعية	النشاط الفنى	النشاط الرياضى	أعداد المعلمين المطلوبة	الفجوة فى أعداد المعلمين	الزيادة السنوية فى أعداد المعلمين
2012	1706	1238	495	248	248	330	330	4596	3396	
2013	1768	1283	513	257	257	342	342	4762	3562	166
2014	1840	1335	534	267	267	356	356	4955	3755	193
2015	1914	1389	556	278	278	370	370	5155	3955	200
2016	1985	1440	576	288	288	384	384	5345	4145	190
2017	2044	1483	593	297	297	396	396	5505	4305	160
2018	2091	1518	607	304	304	405	405	5633	4433	128
2019	2134	1549	620	310	310	413	413	5749	4549	116
2020	2176	1579	632	316	316	421	421	5860	4660	111
2021	2213	1606	642	321	321	428	428	5960	4760	100
2022	2251	1634	654	327	327	436	436	6063	4863	103
2023	2288	1661	664	332	332	443	443	6163	4963	100
2024	2325	1687	675	337	337	450	450	6261	5061	98
2025	2360	1713	685	343	343	457	457	6357	5157	96
2026	2395	1739	695	348	348	464	464	6452	5252	95

أعداد المعلمين الفعلية ٢٠١١	1200	

اهمية تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية فى الحد من ألازمات والكوارث

(دراسة تحليلية على إحدى شركات الإتصالات)

عصام عطية عبدالمنعم

الملخص

تعد إدارة السلامة والصحة المهنية وفي ظل العديد من الازمات والكوارث فى الاونة الاخيرة من المداخل العلمية الحديثة التي تمارسها المنظمات الانتاجية والخدمية على حد السواء لتوفير بيئة عمل آمنة من المخاطر المختلفة ورفع مستوى كفاءة وسائل الوقاية، ويهدف البحث الى تحليل العلاقة بين ادارة السلامة والصحة المهنية والأزمات والكوارث ودور إدارة السلامة والصحة المهنية في الحد من الأزمات والكوارث وكلما استطاعت الشركة المصرية للاتصالات الساعية الى الاستمرار والنمو في ظل العديد من الازمات والكوارث فى الاونة الاخيرة وفي ظل المنافسة المتزايدة من تطبيق قواعد وإجراءات السلامة والصحة المهنية كلما أدى ذلك الى مواجهة الازمات والكوارث التي قد تجابهها بالاسلوب الذي يمكنها من معالجتها وأيجاد السبل الكفيلة لتفاديها مستقبلاً عند تكرارها خاصة وان الظروف الاقتصادية التي تمر بها الشركة المصرية للاتصالات تتطلب المزيد من التحسين والانتباه الى مصادر الازمات والكوارث وادارتها بالاسلوب العلمي المناسب لتجاوزها دون اضرار ناهيك عن الخسائر المادية والمعنوية التي قد تتعرض لها الشركة جراء هذه الازمات . وقد توصلت الدراسة الى عدة استنتاجات اعتمدت عليها التوصيات المناسبة.

الكلمات الدالة

(التحليل الوصفى - التحليل العاملى - عينة طبقية - معاملات الارتباطات البسيطة - معاملات الانحدار القياسية)

مقدمة البحث

لقد أدى التطور التقني الذي شهده العالم إلى ظهور العديد من المخاطر والازمات و التي ينبغي على الإنسان إدراكها وتجنب الوقوع في مسبباتها ، فأماكن العمل المتعددة والمختلفة من ورش ومصانع ومختبرات ومعامل تعتبر بيئات عمل تكثر فيها العديد من المخاطر المهنية التي من الممكن ان تتسبب في حدوث العديد من الازمات والكوارث في بيئة العمل. (موقع دليل السلامة والصحة المهنية – http://www.education.gov.bh)

لذلك فإن توفير بيئة عمل آمنة من مخاطر الصناعات المختلفة ورفع مستوى كفاءة وسائل الوقاية سيؤدي بلا شك إلى الحد من الإصابات والإمراض المهنية ، وحماية العاملين من الحوادث والتى قد تتسبب فى حدوث العديد من الازمات والكوارث .

(المركز العالمي للصحة والسلامة المهنية - http://www.salama-libya.org)

وفي ظل التأثير السلبي الذي تخلفه الأمراض والإصابات المرتبطة بالعمل على صحة العمال وإنتاجيتهم وبالتالي على عائلاتهم والوضع الاقتصادي والرفاهية الاجتماعية في البلاد، ازدادت التوعية حول الصحة والسلامة المهنية والازمات والكوارث في العالم بأكمله. (خوان سومافيا ، 2005).
وما زالت الأمراض والحوادث المهنية أهم أسباب الإصابات والوفيات للعاملين في بيئة العمل ويمكن تفادي أغلب الحوادث والازمات والكوارث عبر "وقاية سليمة تدعمها ممارسات ملائمة للتفتيش والتقرير وترشدها اتفاقيات منظمة العمل الدولية". (منظمة العمل الدولية ، جنيف، 2007)

ومن هنا زاد الاهتمام بالسلامة والصحة المهنية وسنت القوانين والتشريعات وانعقدت المؤتمرات المحلية والدولية المتلاحقة الهادفة لحماية الإنسان في العمل مثل الاتفاقية الدولية (اتفاقية السلامة والصحة والمهنية) رقم (155) لعام 1981 والتي تهدف إلى الوقاية من الحوادث والإصابات الصحية الناجمة عن العمل والتقليل من المخاطر المرتبطة ببيئة العمل بالإضافة إلى قانون العمل المصري رقم (12) الصادر لسنة (2003). ومن أجل تمكين الشركات من تحقيق أهدافها والقيام بدورها لابد من توفير إجراءات السلامة في الشركات وإيجاد بيئة عمل خالية من المخاطر والازمات وإلى تعريف العاملين بقواعد السلامة بهدف الوقاية من مخاطر العمل وتحقيق أكبر قدر من الصحة المهنية وأقل قدر من الخسائر المادية والبشرية وهذا يساعد على زيادة الإنتاج ويخفض التكلفة (أبوعبدون،2003).

لذلك فإن هذه الدراسة سوف تسلط الضوء على دور إدارة السلامة والصحة المهنية في الحد من الازمات والكوارث.

مشكلة البحث

تعتبر المصرية للاتصالات شركة رائدة في مجال الاتصالات السلكية واللاسلكية ونقل المعلومات ذات السرعات الفائقة وهي اعرق شركات الاتصالات في افريقيا والشرق الاوسط .

إلا إن القصور في تدريب وتوعية العاملين بإجراءات ومتطلبات السلامة والصحة المهنية في هذه الشركة وعدم مطابقتها بالمعايير الدولية الموضوعة من قبل منظمة العمل الدولية (ILO) وإدارة السلامة والصحة المهنية (OSHA) ومنظمة الصحة العالمية (WHO) يؤدي ذلك إلى وقوع بعض الحوادث والازمات وقد تشمل الشركة نفسها أو قد تمتد إلى خارجها (المديفر، 2005) لذا يجب أن يتوفر الجهد لرفع مستوى الوعي لدى كل من العاملين والادارة بإهمية السلامة والصحة المهنية في مواقع العمل بالشركة، كما لا بد من التركيز على مجال التدريب والاستفادة من برامج السلامة الشاملة (كوهن وآخرون، 1995) ،

مما تقدم فان مشكلة البحث تبرز من خلال ما يلى :-

- هل هناك ارتباط بين تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية والازمات والكوارث في الشركة المصرية للاتصالات؟

أهمية البحث

إنبثقت أهمية الدراسة من تحديد الآثار الإيجابية من تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية في الحد من الأزمات و الكوارث من خلال تطبيق منظومة ادارة السلامة والصحة المهنية بكفاءة و فاعلية مما تساعد في مقاومة الأزمات و الكوارث و تعجل من سرعة إحتوائها و إعادة الأمن و الإستقرار و إزالة آثارها السلبية و كذلك دراسة الأزمات و الكوارث السابقة بما يكفل فاعلية مواجهتها و إحباط آثارها مستقبلاً ،

وإن عملية تقدير القيمة الاقتصادية للعنصر البشري في أي وحدة إنتاجية يجب أن تنال اهتماماً كبيراً وأن يكون هناك أسلوب دقيق يجب إتباعه واتساع دائرة تطبيقه في جميع الشركات للخروج بصورة دقيقة عن قيمة الثروة البشرية ومدى العائد من استثماره في التدريب والتعليم والخبرة حيث توضح أهمية الدراسة فيما يلي :

- تساهم هذه الدراسة للباحث في إثراء معلوماته حول السلامة والصحة المهنية ، وكيفية تجنب المخاطر المهنية والازمات والوقاية منها .

- تسهم الدراسة في معرفة العاملين بإشتراطات السلامة والصحة المهنية عند أدائهم لأعمالهم وبالتالي التقليل من حجم الخسائر سواء المادية أو البشرية في الشركة المصرية للاتصالات

- تساعد هذه الدراسة العاملين بالشركات في الأخذ بكافة احتياطات السلامة المهنية والقواعد والإجراءات الواجب إتباعها في بيئة العمل ، وذلك من أجل خلق جو آمن وبيئة خالية من المخاطر والازمات في جميع المجالات.
- توضح الدراسة مدى احتياج إدارة الشركة المصرية للاتصالات لنتائج هذه الدراسة والتي من شأنها القيام بتطوير قواعد وإجراءات السلامة والصحة المهنية واساليب إدارة الازمات والكوارث .
- تساعد نتائج هذه الدراسة المسئولين عن الشركة المصرية للاتصالات في تفعيل قواعد وإجراءات السلامة والصحة المهنية عند قيامهم بعملهم في الشركة.

أهداف البحث

أن الهدف الرئيسى من هذا البحث هو ابراز اهمية دور تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية بالشركة المصرية للإتصالات في الحد من الازمات والكوارث وذلك من خلال قياس المتغيرات المستقلة ، والمتغيرات التابعة ، وذلك ببعض العبارات التى تناسب كل متغير.

منهج البحث وخطواته

سوف يتم إستخدام المنهج الوصفى التحليلى عن طريق مدخل المسح بإستخدام الإستبانة كأداة لجمع البيانات بالاضافة الى اساليب اخرى وفيما يلى توضيحها .

جمع البيانات

سوف يعتمد اسلوب العمل في أنجاز الجانب النظري من الدراسة الأسلوب الوصفي باستخدام المراجع العلمية والدوريات وأدبيات إدارة الإنتاج والعمليات فضلا عن شبكة الانترنت ومحتوياتها الحديثة. اما الجانب الميداني من الدراسة سوف يعتمد بشكل رئيسي على استمارات الاستبيان التي جاءت متوافقة ومستمدة من الجانب النظري من الدراسة بما يمكن تحديد أهمية تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية في الحد من الازمات والكوارث بالشركة المصرية للإتصالات اذ وزعت (275) مائتان وخمسة وسبعون استمارة على مسؤولي الوحدات ورؤساء الاقسام زيادة على ذلك المقابلات الشخصية والزيارات الميدانية للادارات ومواقع الشركة المختلفة بالمحافظات المختلفة . كما اعتمد اسلوب ليكرت الخماسى على تحليل الاستمارة وفضلآ عن الارتباط بين متغييرات الدراسة الى جانب التكرارات والنسب المئوية والمتوسطات والانحرافات المعيارية ومعامل الاختلاف لوصف وتشخيص متغيرات الدراسة

مجتمع الدراسة :

تقوم الشركة المصرية للاتصالات بدور حيوى وهام فى خدمة المجتمع من خلال توفير خدمة الانترنت فائق السرعة بما ساهم فى سهولة تدوال البيانات بما يرفع من الدخل القوى للبلاد .

وقد قام الباحث بسحب عينة طبقية من مختلف المستويات الوظيفية بالادارة العامة للسلامة والصحة المهنية وجميع الاقسام التابعة لها فى قطاعات الشركة المختلفة بالمحافظات وكذلك الادارة العامة لإدارة وتحليل الازمات وجميع الاقسام التابعة لها فى قطاعات الشركة المختلفة بالمحافظات ومبنى سنترال القبة وحيث أن مجتمع الدراسة المتمثل فى الادارة العامة للسلامة والصحة المهنية وجميع الاقسام التابعة لها فى قطاعات الشركة المصرية للاتصالات المختلفة بالمحافظات والبالغ عدد مفردات الدراسة من واقع سجلات شئون العاملين وهم (90) مفردة وكذلك الادارة العامة لإدارة وتحليل الازمات وجميع الاقسام التابعة لها فى قطاعات الشركة المصرية للاتصالات المختلفة بالمحافظات والبالغ عدد مفردات الدراسة من واقع سجلات شئون العاملين وهم(50) مفردة .

وكذلك إدارة مبنى سنترال القبة والبالغ عدد مفردات الدراسة من واقع سجلات شئون العاملين وهم(360) مفردة . وبالتالى يكون اجمالى مفردات مجتمع الدراسة البالغ (500)مفردة مقسمة كما فى جدول رقم (1) .

جدول رقم (1)

توزيع مفردات المجتمع على الإدارات

البيـــــــان	العدد	النسبة من المجتمع (%)
إدارة مبنى سنترال القبة	360	72
الإدارة العامة للسلامة والصحة المهنية وجميع الاقسام التابعة لها فى الفروع	90	18
الادارة العامة لإدارة وتحليل الازمات وجميع الاقسام التابعة لها فى الفروع	50	10
المجموع	500	100

ويتضح من الجدول رقم (1) توزيع مفردات مجتمع الدراسة على الإدارات .

4-عينة الدراسة :

وتم سحب عينة طبقية تمثل (55%)من مجتمع الدراسة اى تمثل(275) مفردة منهم (60) مفردة من الادارة العامة للسلامة وجميع الاقسام التابعة لها ومنهم ايضا (35) مفردة من الادارة العامة لإدارة وتحليل الازمات وايضا (180) مفردة من العاملين بإدارة مبنى سنترال القبة كما فى جدول رقم (9) .

وسوف يتم نشر(275) إستمارة إستقصاء صالحة للتحليل الاحصائى وذلك نظراً لضيق الوقت ولعدم إمكانية نشر (500) إستمارة وإعادة جمعهم وتحليلهم وإستخراج النتائج والتوصيات

جدول رقم (2)

توزيع مفردات العينة على الإدارات

البيـــــــان	العدد	النسبة من العينة (%)	النسبة من مفردات المجتمع (%)
إدارة مبنى سنترال القبة	180	65	50
الأدارة العامة للسلامة والصحة المهنية وجميع الاقسام التابعة لها فى الفروع	60	22	67
الادارة العامة لإدارة وتحليل الازمات وجميع الاقسام التابعة لها فى الفروع	35	13	70
المجموع	275	100	------

ويتضح من الجدول رقم (2) توزيع مفردات عينة الدراسة على الإدارات ونسبتها من مفردات المجتمع

متغيرات الدراسة

حيث تلاحظ ان متغيرات الدراسة متمثلة فى الاستبيان والذى يوضح مجموعتين من المتغيرات وهما متغيرات المجموعة الأولى (X) الخاصة بتدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية

ومتغير المجموعة الثانية (Y) وهى تتكون من مؤشر خاص بفاعلية تطبيق اجراءات السلامة و الصحة المهنية فى الحد من الازمات والكوارث

وفضلا عن ايضاحها فى استمارة الاستبيان وكيفية قياس كل متغير فيها.

الاسلوب البحثى

أستخدم الأسلوب التحاليلي في الدراسة حيث استخدم التحليل الوصفى من خلال الوسط الحسابى والانحراف المعيارى ومعامل الاختلاف واستخدم التحليل العاملي Factor Analysis لاستخلاص المتغيرات المستقلة والتابعة من عناصر صحيفة الأستبيان ثم تم حساب معاملات الأرتباط البسيطة في شكل مصفوفة الارتباطات لتحديد شدة الارتباط بين المتغيرات المستقلة (تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية) والمتغير التابع (فاعلية تطبيق اجراءات السلامة و الصحة المهنية فى الحد من الازمات والكوارث).

نتائج الدراسة

هذا الجزء يستعرض اهم النتائج التي توصل اليها البحث وتتماشا مع الأهداف البحثية

جدول رقم (3)

الأوزان الناتجة من التحليل العاملي

المتغيرات (Variables)	الاوزان القياسية	العوامل (Factors)
• يتلقى العاملون تدريباً على خطة الإخلاء والطوارئ ومكافحة الحريق • يتم اختبار فاعلية خطط الطوارىء وإدارة الازمات والكوارث الصناعية والطبيعية واجراء تدريبات عملية عليها للتأكد من كفاءتها بصفة دورية • يتم التدريب على طرق التعامل مع أنظمة وأجهزة السلامة بالسنترال • هناك تدريب للعاملين على الإسعافات الأولية وكيفية استخدام مهمات الوقاية الشخصية ايضا • يتم التدريب على إجراءات السلامة عن طريق(التدريب العملي-المحاضرات - النشرات المطبوعة- ورش العمل) • يتم توعية وتدريب العاملين بالمخاطر التى يواجهونها والزامهم بإستخدام وسائل الوقاية المقررة • هناك مشاركة للعاملين في الندوات والمؤتمرات الدولية المتعلقة بالسلامة والصحة المهنية) (منظمة العمل الدولية –منظمة الصحة العالمية) • تلقيت تدريب على إجراءات السلامة من قبل جهات خارجية (الدفاع المدني – وزارة الصحة –وزارة العمل –مؤسسات أهلية)	.584 .530 .506 .292 .581 .725 .421 .272	تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية ← (X)
• تطبيق نظام ادارة السلامة والصحة المهنية ادى الى تأمين بيئة العمل والحد من الحوادث والامراض • تطبيق نظام ادارة السلامة والصحة المهنية ادى الى رفع الروح المعنوية وارضاء العاملين • تطبيق نظام ادارة السلامة والصحة المهنية ادى الى زيادة الانتاجية للعاملين • تطبيق نظام ادارة السلامة والصحة المهنية ادى الى زيادة جودة العمل • تطبيق نظام ادارة السلامة والصحة المهنية ادى الى الحماية من اى عقوبات قانونية يفرضها القانون طبقا لقانون العمل • تطبيق نظام ادارة السلامة والصحة المهنية ادى الى خفض التكاليف والنفقات المباشرة والغير مباشرة وبالتالى زيادة فى الارباح	.772 .843 .686 .168 -.116 -.081	فاعلية تطبيق اجراءات السلامة والصحة المهنية فى الحد من الازمات والكوارث← (Y)

ويتضح من جدول رقم (3) الأوزان الناتجة من التحليل العاملي ،

و كانت اهم النتائج تمثلت فى:

- أعلى الاوزان القياسية فى مؤشر تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية (X (هو 725. والتى تمثل بعبارة يتم توعية وتدريب العاملين بالمخاطر التى يواجهونها والزامهم بإستخدام وسائل الوقاية المقررة

- أعلى الاوزان القياسية فى مؤشر فاعلية تطبيق اجراءات السلامة والصحة المهنية فى الحد من الازمات والكوارث(Y)هو 843. والتى تمثل بعبارة تطبيق نظام ادارة السلامة والصحة المهنية ادى الى رفع الروح المعنوية وارضاء العاملين وذلك لما لها بالغ الاثر والاهمية فى الحد من الازمات والكوارث

جدول رقم (4)

الوسط الحسابى ومعامل الاختلاف لمؤشرات تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية

العبـــــــــــارة	الوسط الحسابى (Mean)	الانحراف المعيارى (Std. Deviation)	معامل الاختلاف (C.V)
يتلقى العاملون تدريباً على خطة الإخلاء والطوارئ ومكافحة الحريق	**4.17**	**.62**	**.15**
يتم اختبار فاعلية خطط الطوارىء وإدارة الازمات والكوارث الصناعية والطبيعية واجراء تدريبات عملية عليها للتأكد من كفاءتها بصفة دورية	**3.99**	**.59**	**.15**
يتم التدريب على طرق التعامل مع أنظمة وأجهزة السلامة بالسنترال	**3.85**	**.54**	**.14**
هناك تدريب للعاملين على الإسعافات الأولية وكيفية استخدام مهمات الوقاية الشخصية ايضا	**4.01**	**.49**	**.12**
يتم التدريب على إجراءات السلامة عن طريق (التدريب العملي–المحاضرات – النشرات المطبوعة –ورش العمل)	**4.27**	**.62**	**.15**
يتم توعية وتدريب العاملين بالمخاطر التى يواجهونها والزامهم بإستخدام وسائل الوقاية المقررة	**4.19**	**.64**	**.15**
هناك مشاركة للعاملين في الندوات والمؤتمرات الدولية المتعلقة بالسلامة والصحة المهنية (منظمة العمل الدولية– منظمة الصحة العالمية)	**3.86**	**.62**	**.16**
تلقيت تدريب على إجراءات السلامة من قبل جهات خارجية (الدفاع المدني– وزارة الصحة– وزارة العمل– مؤسسات أهلية)	**3.68**	**1.04**	**.28**

ويتضح من الجدول السابق رقم (4) ان متوسط المجموع للوسط الحسابى لمؤشرات تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية ما يقرب من (4.00)وهذا ان دل فأنه يدل على قبولها والموافقة عليها من مفرادات عينة الدراسة وان معامل الاختلاف لمؤشرات تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية ما يقترب من بعضها البعض وهذا ان دل فأنه يدل على ثبات فى الرأى من مفردات عينة الدراسة وموافقتهم على هذه المؤشرات

جدول رقم (5)

نسبة الموافقين من مفردات عينة الدراسة لمؤشرات تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية

العبـــــــــارة	Frequencyالتكرار	Percentالنسبة
يتلقى العاملون تدريباً على خطة الإخلاء والطوارئ ومكافحة الحريق	**161**	**58.5**
يتم اختبار فاعلية خطط الطوارىء وإدارة الازمات والكوارث الصناعية والطبيعية واجراء تدريبات عملية عليها للتأكد من كفاءتها بصفة دورية	**180**	**65.5**
يتم التدريب على طرق التعامل مع أنظمة وأجهزة السلامة بالسنترال	**188**	**68.4**
هناك تدريب للعاملين على الإسعافات الأولية وكيفية استخدام مهمات الوقاية الشخصية ايضا	**208**	**75.6**
يتم التدريب على إجراءات السلامة عن طريق (التدريب العملي– المحاضرات – النشرات المطبوعة –ورش العمل)	**149**	**54.2**
يتم توعية وتدريب العاملين بالمخاطر التى يواجهونها والزامهم بإستخدام وسائل الوقاية المقررة	**154**	**56.0**
هناك مشاركة للعاملين في الندوات والمؤتمرات الدولية المتعلقة بالسلامة والصحة المهنية (منظمة العمل الدولية– منظمة الصحة العالمية)	**189**	**68.7**
تلقيت تدريب على إجراءات السلامة من قبل جهات خارجية (الدفاع المدني– وزارة الصحة– وزارة العمل– مؤسسات أهلية)	**156**	**56.7**

ويتضح من الجدول السابق رقم (5) ان نسبة الموافقين من مفرادات عينة الدراسة لمؤشرات تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية اكبر من 50% وهذا ان دل فأنه يدل على ان مؤشرات تدريب وتوعية العاملين بإجراءات السلامة والصحة المهنية تم قبولها والموافقة عليها من اغلب مفرادات عينة الدراسة .

جدول رقم (6)

الوسط الحسابى ومعامل الاختلاف لمؤشرات فاعلية تطبيق نظام ادارة السلامة والصحة المهنية فى الحد من الازمات والكوارث

العبـــــــــارة	الوسط الحسابى (Mean)	الانحراف المعيارى (Std. Deviation)	معامل الاختلاف (C.V)
تطبيق اجراءات السلامة والصحة المهنية ادى الى تأمين بيئة العمل والحد من الحوادث والامراض	4.54	.50	.11
تطبيق اجراءات السلامة والصحة المهنية ادى الى رفع الروح المعنوية وارضاء العاملين	4.17	.76	.18
تطبيق اجراءات السلامة والصحة المهنية ادى الى زيادة الانتاجية للعاملين	4.16	.44	.11
تطبيق اجراءات السلامة والصحة المهنية ادى الى زيادة جودة العمل	4.13	.43	.10
تطبيق اجراءات السلامة والصحة المهنية ادى الى الحماية من اى عقوبات قانونية يفرضها القانون طبقا لقانون العمل	4.23	.56	.13
تطبيق اجراءات السلامة والصحة المهنية ادى الى خفض التكاليف والنفقات المباشرة والغير مباشرة وبالتالى زيادة فى الارباح	4.40	.55	.13

ويتضح من الجدول السابق رقم (6) ان متوسط المجموع للوسط الحسابى لمؤشرات فاعلية تطبيق اجراءات السلامة والصحة المهنية فى الحد من الازمات والكوارث (4.27) وهذا ان دل فأنه يدل على قبولها والموافقة عليها من مفرادات عينة الدراسة والاتفاق على ان تطبيق اجراءات السلامة والصحة المهنية بفاعلية يؤدى الى الحد من الازمات والكوارث وان معامل الاختلاف لمؤشرات فاعلية تطبيق اجراءات السلامة والصحة المهنية فى الحد من الازمات والكوارث ما يقترب من بعضها البعض وهذا ان دل فأنه يدل على ثبات فى الرأى من اغلب مفردات عينة الدراسة وحيث تلاحظ ان معامل الاختلاف لمؤشر تطبيق اجراءات السلامة والصحة المهنية ادى الى رفع الروح المعنوية وارضاء العاملين **(18.)** اعلى قيمة معامل الاختلاف ولذلك يجب ان تنال اكثر اهتماما من الادارة العليا بالشركة لما لها بالغ الاثر والاهمية فى فاعلية تطبيق اجراءات السلامة والصحة المهنية فى الحد من الازمات والكوارث بالشركة

جدول رقم (7)

نسبة الموافقين من مفردات عينة الدراسة لمؤشرات فاعلية تطبيق نظام ادارة السلامة والصحة المهنية فى الحد من الازمات والكوارث

العبـــــــــارة	Frequencyالتكرار	Percentالنسبة
تطبيق اجراءات السلامة والصحة المهنية ادى الى تأمين بيئة العمل والحد من الحوادث والامراض	126	45.8
تطبيق اجراءات السلامة والصحة المهنية ادى الى رفع الروح المعنوية وارضاء العاملين	109	39.6
تطبيق اجراءات السلامة والصحة المهنية ادى الى زيادة الانتاجية للعاملين	215	78.2
تطبيق اجراءات السلامة والصحة المهنية ادى الى زيادة جودة العمل	219	79.6
تطبيق اجراءات السلامة والصحة المهنية ادى الى الحماية من اى عقوبات قانونية يفرضها القانون طبقا لقانون العمل	174	63.3
تطبيق اجراءات السلامة والصحة المهنية ادى الى خفض التكاليف والنفقات المباشرة والغير مباشرة وبالتالى زيادة فى الارباح	148	53.8

ويتضح من الجدول السابق رقم (7) ان نسبة الموافقين من مفرادات عينة الدراسة لمؤشرات فاعلية تطبيق نظام ادارة السلامة والصحة المهنية فى الحد من الازمات والكوارث اكبر من 50 % ماعدا مؤشر تطبيق نظام ادارة السلامة والصحة المهنية ادى الى رفع الروح المعنوية وارضاء العاملين **39.6** % ومؤشر تطبيق نظام ادارة السلامة والصحة المهنية ادى الى تأمين بيئة العمل والحد من الحوادث والامراض **45.8** % ولذلك يجب ان يأخذوا اكثر اهتماما من الادارة العليا بالشركة لما لهم بالغ الاثر والاهمية فى مؤشرات فاعلية تطبيق نظام ادارة السلامة والصحة المهنية فى الحد من الازمات والكوارث بالشركة وقبولهم والموافقة عليهم من جميع مفرادات عينة الدراسة .

جدول رقم (8)

مصفوفة الأرتباطات البسيطة بين متغيرات الدراسة

		تدريب وتوعية العاملين على إجراءات السلامة والصحة المهنية (X)	فاعلية تطبيق اجراءات السلامة والصحة المهنية فى الحد من الازمات والكوارث (Y)
تدريب وتوعية العاملين على إجراءات السلامة والصحة المهنية (X)	Pearson Correlation	1	.339**
	Sig. (2-tailed)		.000
	Sig. (2-tailed)	.000	.119
فاعلية تطبيق إجراءات السلامة والصحة المهنية فى الحد من الازمات والكوارث (Y)	Pearson Correlation	.339**	1
	Sig. (2-tailed)	.000	

ويتضح من جدول رقم (8) انه يوجد ارتباط بين تدريب وتوعية العاملين على إجراءات السلامة والصحة والمتمثل فى مؤشر (X) و فاعلية تطبيق إجراءات السلامة والصحة المهنية فى الحد من الازمات والكوارث التابع لإدارة الازمات والكوارث والمتمثل فى مؤشر(Y)

وكانت اهم النتائج تمثلت فى :

فاعلية تطبيق اجراءات السلامة والصحة المهنية فى الحد من الازمات والكوارث ترتبط ارتباط معنوى مع تدريب وتوعية العاملين على إجراءات السلامة والصحة وذلك نظرا لآن Sig اقل من 05.

جدول رقم (9)
معامل الانحدار القياسى

المتغيرات	تدريب وتوعية العاملين على إجراءات السلامة والصحة المهنية (X)
فاعلية تطبيق إجراءات السلامة والصحة المهنية فى الحد من الازمات والكوارث (Y)	.23

ويتضح من جدول رقم (9) مؤشر تدريب وتوعية العاملين على إجراءات السلامة والصحة المهنية (X) واثره على مؤشر فاعلية تطبيق إجراءات السلامة والصحة المهنية فى الحد من الازمات والكوارث وذلك بإستخدام قيم Beta Coefficient حيث بلغت 23% وهذا ان دل فأنه يدل على وجود علاقة قوية بينهم واثر مؤشر تدريب وتوعية العاملين على إجراءات السلامة والصحة المهنية على مؤشر فاعلية تطبيق إجراءات السلامة والصحة المهنية فى الحد من الازمات والكوارث

خلاصة الدراسة

- ويتضح ان تدريب وتوعية العاملين على إجراءات السلامة والصحة المهنية لها بالغ الأثر والاهمية فى الحد من الازمات والكوارث وان القصور من قبل إدارة الشركة فى تدريب وتوعية العاملين على إجراءات السلامة والصحة المهنية يؤدى الى حدوث العديد من الازمات والكوارث والتى يجب ان تنال العديد من الاهتمال من قبل إدارة الشركة .
- ويتضح أن تدريب وتوعية العاملين على إجراءات السلامة والصحة المهنية فى الشركة أدى الى رفع الروح المعنوية وارضاء العاملين أولا ثم يليها تأمين بيئة العمل والحد من الحوادث والامراض ثم يليها زيادة الانتاجية للعاملين ثم يليها زيادة جودة العمل ثم يليها الحماية من اى عقوبات قانونية يفرضها القانون طبقا لقانون العمل ثم يليها خفض التكاليف والنفقات المباشرة والغير مباشرة وبالتالى زيادة فى الارباح وذلك بحساب قيم الاوزان القياسية وان النتائج السابقة لها بالغ الاثر والاهمية فى الحد من الازمات والكوارث التى من الممكن ان تمر بها الشركة .

توصيات الدراسة

- لا بد من الاهتمام بعنصر التدريب والمعرفة لدى العاملين، بما يضمن لهم الحماية اللازمة من مخاطر العمل ووضع وتنفيذ برامج التدريب المستمر بهدف تطوير خبراتهم ومهاراتهم التقنية وتنمية الوعي الصحي لديهم.
- التأكيد على التزام الإدارة العليا بالاهتمام بصفة شخصية بأنشطة وخطط وبرامج السلامة وإعطاء مسألة السلامة ضمن الأولويات في اجتماعات مجلس الادارة .
- التعاون والتنسيق المستمر مع الجهات الخارجية (الجهات الحكومية – المؤسسات الأهلية) لأن نجاح هذه المسألة يتطلب وعياً عميقاً منهم بالمسؤولية المشتركة يدفع لبذل الجهد الطيب في هذا المجال.
- إجراء المزيد من الدراسات والأبحاث حول موضوع السلامة والصحة المهنية من أجل الوصول إلى نتائج أكثر عمقاً تساهم في تطور قطاعات الشركة بشكل أفضل.
- السعي نحو الحصول على شهادة **IS045001** وشهادة **ISO 14001**، لضمان السلامة والصحة المهنية والبيئة بالشركة كخطوة جديدة نحو تطبيق نظام الإدارة المتكامل **(IMS) (Integrated Management System)**

المراجع العربية

م	العربية
1	أبو شيخه، نادر (2010) : إدارة الموارد البشرية ، الطبعة الثانية ، دار الصفاء للنشر والتوزيع ، عمان ، الأردن
2	العقدة، السيد (2018) : الامراض المهنية واثرها على الصحة والسلامة المهنية ، المؤتمر الدولى الثالث للسلامة والصحة المهنية والمنعقد فى الفترة من 5-7 مارس ، القاهرة ، اكاديمية المستقبل
3	الزدجالى، اسماعيل (2018) : معايير العاملين فى قطاع الصحة والسلامة المهنية ، المؤتمر الدولى الثالث للسلامة والصحة المهنية والمنعقد فى الفترة من 5-7 مارس ، القاهرة ، اكاديمية المستقبل
4	الرفاعى ، ممدوح (2011) : الفساد الادارى والبيئى والمجتمع ، المؤتمر السنوى السادس عشر للازمات والكوارث ،القاهرة، كلية التجارة، جامعة عين شمس.
5	العباسى ، عبدالحميد ، (2016) " تحليل البيانات بإستخدام الحزم الاحصائية " معهد الدراسات والبحوث الاحصائية ، جامعة القاهرة .
6	الروسان وآخرون (2009) : الأمن الصناعي والسلامة المهنية ، الطبعة الثانية ، مكتبة المجتمع العربي للنشر والتوزيع، عمان.
7	الشيخ خليل، علي (2008) : تقييم وسائل الوقاية والسلامة المستخدمة في مستشفيات قطاع غزة الحكومية وأثرها على أداء العاملين ، (رسالة ماجستير) ، الجامعة الإسلامية، غزة.
8	الروسان وآخرون (2009) : الأمن الصناعي والسلامة المهنية ، الطبعة الثانية ، مكتبة المجتمع العربي للنشر والتوزيع، عمان.
9	قطيشات، تالا وآخرون (2007) : مبادئ في الصحة والسلامة العامة ، الطبعة الثانية ، دار المسيرة للنشر والتوزيع ، عمان ، الأردن .
10	المغني، أميمه (2006) : واقع وإجراءات الأمن والسلامة المهنية المستخدمة في منشآت قطاع الصناعات التحويلية في قطاع غزة ، (رسالة ماجستير) ، الجامعة الإسلامية ، غزة.
11	المديفر، فهد (2005) : مدى فعالية تطبيق أنظمة الأمن والسلامة المهنية والتقنية ، دراسة مسحية على معامل الأقسام العلمية بكليات البنات ، الرياض (رسالة ماجستير) ، جامعة نايف العربية للعلوم الأمنية ، الرياض.
12	برى ، عدنان و هندى ، محمود (2009) : " مبادىء الاحصاء والاحتمالات " ، مكتبة الشقرى للنشر والتوزيع
13	طبية ، احمد عبدالسميع (2008) : " مبادىء الاحصاء " ، دار البداية ، عمان ، ط1
14	وزارة القوى العاملة (2003) : قانون العمل المصري رقم (12) لعام (2003) ،

المراجع الأجنبية

15 Allen,D.B,J.H.Burlon and J.D.Hott (1983) J.Anim.Sci,57:765

16 BS OHSAS 18001, Occupational health and safety management systems – Requirements, BSI Group, 2009

17 Cristine Person and Ian Mitroff: " From Crisis Prone To Crisis Prepared: A Framework For Crisis Management", Academy of Management Excutive . Vol 7, No 1, 1993

18 Davis,C.D.(2002)statistical methods for the Analysis of Repeate.1 Measurements. Springer.veriag,New York.

19 Dejoy , D., Schaffer , B. & Wilson, N. , (2003) : " Creating Safer Workplaces : assessing the determinants and role of Safety Climate " , Journal of Safety Research , USA , 2003.

20 Graham Allison: Essence of Decision, U.S.A., Little Brown and Company, 1971, P.5.

21 Hinke lmann, K. and O. kempthorn (2005) . Design and analysis of Experiments: Advanced Experimental Design and analysis of Experiments: Advanced Experimental Design.vol.2.John wiley&Sons, New York.

22 ISO 45001 Occupational Health and Safety Management System – Draft Standard, International Standards Organization, 2016

المواقع الالكترونية

23 http//:www.ilo.org/global

24 http://www.alolabor.org

25 http://www.education.gov.bh -

26 http://www.safety-eng.com -

27 http://www.salama-libya.org -