Box-Jenkins Methodology to TS Modeling

Definition

The Box-Jenkins methodology is a systematic approach for identifying, estimating, and forecasting Autoregressive Integrated Moving Average (ARIMA) models. It is widely used in forecasting time series data. It consists of three main stages:

1. Model Identification
2. Parameter Estimation
3. Model Diagnostics & Forecasting

ARIMA(p, d, q) Modelling Structure

An ARIMA model is defined by three parameters:

1. p: Order of the **Autoregressive (AR)** component
2. d: Degree of **differencing**
3. q: Order of the **Moving Average (MA)** component

An ARIMA model combines:

1. **AR(p)**:
$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \epsilon_t$$
2. **I(d)**: Differencing to achieve stationarity
3. **MA(q)**:
$$X_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$$

Therefore, the general form of an **ARIMA(p, d, q)** model is:

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)(1 - B)^d X_t = c + (1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q)\varepsilon_t$$

Steps in Box-Jenkins Methodology

Step 1: Model Identification

1. Check for stationarity: Use plots and statistical tests (e.g., ADF test)
2. Differencing: Apply differencing to achieve stationarity
3. Plot ACF and PACF: Identify potential AR and MA orders

Step 2: Model Estimation

1. Estimate ARIMA parameters using methods such as maximum likelihood or least squares

Step 3: Diagnostic Checking

1. Residual analysis: Check if residuals resemble white noise
2. Ljung-Box Q-test: Test residual autocorrelations
3. Jarque-Bera and Sharpiro-Wilk Tests: To test for normality of errors
4. Goodness-of-fit metric: AIC, BIC, RMSE, MAE, MAPE

Step 4: Forecasting

1. Use the fitted ARIMA model to predict future values

Tools for Box-Jenkins Modelling

1. R packages: forecast, tseries, stats
2. Python: statsmodels.tsa.arima.model

Strengths of Box-Jenkins

1. Systematic approach
2. Handles various patterns: trend, seasonality, autocorrelation
3. Well-suited for short-term forecasting

Limitations

1. Assumes linearity
2. Assumes normality of errors
3. Can be complex for beginners

Example 1

Given the following monthly sales data:

Month	Sales
Jan	112
Feb	118
Mar	132
Apr	129
May	121
Jun	135
Jul	148
Aug	148
Sep	136
Oct	119
Nov	104
Dec	118

Apply the Box-Jenkins methodology to obtain a time series model for the data

Step 1: Model Identification

1. Plot Time Series: Observe trend/seasonality
2. ADF Test: Check for stationarity. Suppose it shows the data is non-stationary.
3. Apply Differencing (d = 1): First difference the series.
4. Plot ACF and PACF:
   (a) ACF: Decays slowly $\Rightarrow$ use differencing
   (b) PACF: Significant spike at lag 1 $\Rightarrow$ AR(1)
5. Proposed Model: ARIMA(1,1,0)

Step 2: Model Estimation

Estimate the model parameters using statistical software (e.g., R, Python).
Suppose the model estimated is:

$$X_t - X_{t-1} = 0.7(X_{t-1} - X_{t-2}) + \epsilon_t$$

Step 3: Diagnostic Checking

1. Plot residuals: No obvious pattern
2. Ljung-Box test: p-value exceeds α, this implies that the residuals resemble white noise
3. Model is adequate ✓

Step 4: Forecasting

Forecast future values using the fitted ARIMA(1,1,0) model. For example:
If the last observed values are:

$$X_{t-1} = 118, \quad X_{t-2} = 104$$

Forecast:

$$X_t = X_{t-1} + 0.7(X_{t-1} - X_{t-2}) = 118 + 0.7(118 - 104) = 118 + 9.8 = 127.8$$

Trial Questions

1. Explain how to identify an AR(2) model from the ACF and PACF plots.
2. The ACF of a time series shows a significant spike at lag 1, but none afterward. The PACF tails off. What ARIMA model is appropriate?
3. Given a non-stationary time series, describe the steps to fit an ARIMA model using the Box-Jenkins approach.

JOMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY

DEPARTMENT OF MATHEMATICS AND ACTUARIAL SCIENCE

STA 2401: TIME SERIES ANALYSIS - CAT I
DATE: FEBRUARY 14, 2025. TIME: 75 MINUTES

INSTRUCTIONS: ATTEMPT ALL QUESTIONS

1. What components of time series are applicable in these situations? (3 Marks)

   (a) A car accident recently occurring on the Thika Highway.... **Irregular**

   (b) Increase in agricultural production due to advances in technology....**Trend**

   (c) A decrease in the price of onions during harvesting time....**Seasonal**

2. Describe the objectives of time series analysis. (3 Marks)

   (a) Descriptive Analysis (Decomposition)
   The goal is to understand and summarize the underlying structure of the data by decomposing it into trend, seasonal, cyclical, and irregular components. It aids in identifying key features of the time series and preparing the data for further analysis.

   (b) Forecasting (Prediction)
   The objective is to build models that capture the temporal dynamics of the data, thereby enabling the prediction of future values. Accurate forecasting is essential for planning, decision-making, and risk management in various fields such as finance, economics, and meteorology.

   (c) Explanation (Inference and Hypothesis Testing)
   This involves using statistical methods to test hypotheses about the data and to understand the underlying stochastic process. It includes determining whether the series is stationary, identifying significant relationships in multivariate settings, and assessing the impact of external interventions.

3. Table 1 shows quarterly sales (in KSh. 10,000.00) of a certain product for 4 years.

   (a) Draw a time series plot of the quarterly sales data (NB: Refer to Figure 1 for more details) (3 Marks)

   (b) Provide 2 comments on the time series plot drawn in (a) (NB: Refer to Figure 1 for more details) (2 Marks)

Table 1: Quarterly Sales Data (Full Grid Table)

Year	Quarter I	Quarter II	Quarter III	Quarter IV
2021	4.8	5.8	6.0	6.3
2022	4.1	5.2	5.6	5.9
2023	6.0	6.8	7.5	8.0
2024	6.5	7.4	7.8	8.4

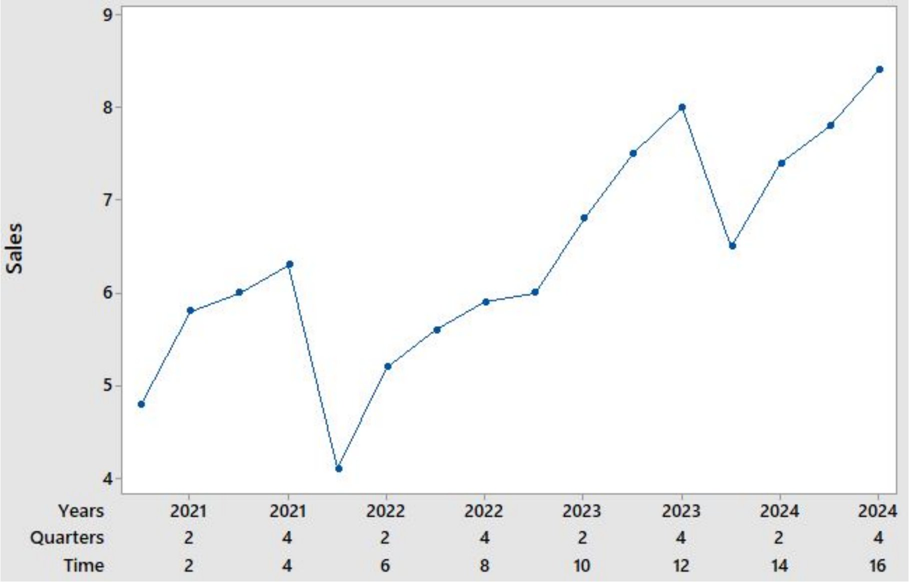

Figure 1: Gas Bills Time Series Analysis Plot (Question 3(a))

(i) The sales show a general upward trend over the four years, indicating growth in demand for the product.

(ii) The trend appeared to be increasing slowly

(iii) There is seasonal variations with sales being highest in the fourth quarter and lowest in the second quarter of every year

(iv) The seasonal variations appeared to be increasing slowly

(v) The components of the time series are changing slowly, and the time series may therefor be described by a multilicative model

(vi) The lowest sale is 4.1 and is recorded in the second quarter of the first year while the highest sales is 8.4 which is recorded in the last quarter of the last year.

(c) Using the information in Table 1, copy and complete Table 3 on page 2 (NB: Refer to Figure 2 for details) (6 Marks)

Years	Time (t)	Quarters	Sales	4yr MT	4 yr MA	2-item MT	4yr CMA
2021	1	1	4.80	*	*	*	*
2021	2	2	5.80	*	*	*	*
				22.90	5.73		
2021	3	3	6.00			11.28	5.64
				22.20	5.55		
2021	4	4	6.30			10.95	5.48
				21.60	5.40		
2022	5	1	4.10			10.70	5.35
				21.20	5.30		
2022	6	2	5.20			10.50	5.25
				20.80	5.20		
2022	7	3	5.60			10.88	5.44
				22.70	5.68		
2022	8	4	5.90			11.75	5.88
				24.30	6.08		
2023	9	1	6.00			12.63	6.31
				26.20	6.55		
2023	10	2	6.80			13.63	6.81
				28.30	7.08		
2023	11	3	7.50			14.28	7.14
				28.80	7.20		
2023	12	4	8.00			14.55	7.28
				29.40	7.35		
2024	13	1	6.50			14.78	7.39
				29.70	7.43		
2024	14	2	7.40			14.95	7.48
				30.10	7.53		
2024	15	3	7.80	*	*	*	*
2024	16	4	8.40	*	*	*	*

(d) Using the same axes, superimpose the *Centered 4-yearly MA* on the time series plot and provide comments on it (NB: Refer to Figure 2 for details) (2 Marks)

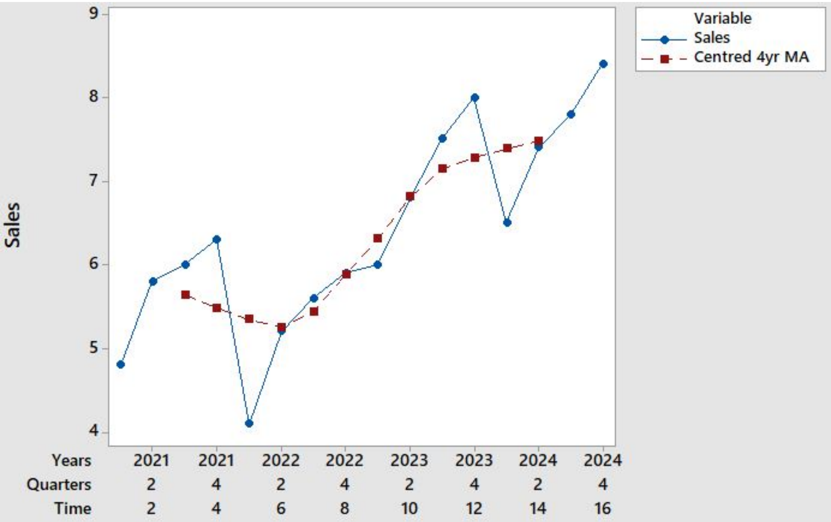

Figure 2: Gas Bills Time Series Analysis Plot

(i) The centered 4-yearly moving average smooths out the seasonal fluctua-tions, revealing the underlying trend more clearly.

(ii) The moving average confirms the upward trend observed in the original time series plot.

(iii) The seasonal variations appeared to be increasing slowly

(iv) There is an increasing linear trend in the data

4. In time series analysis, differentiate weak stationarity from strict stationarity (4 marks)

A time series $\{X_t\}$ is said to be **strictly stationary** if the joint distribution of

$$(X_{t_1}, X_{t_2}, \ldots, X_{t_k})$$

is identical to the joint distribution of

$$(X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_k+h})$$

for all integers $t_1, t_2, \ldots, t_k$, for any $k \geq 1$, and for all time shifts h. This means that every aspect of the distribution (not just the moments) remains invariant under a time shift. In contrast, a time series is said to be **weakly stationary** (or second-order stationary) if only the first two moments are invariant with time. Specifically, a series $\{X_t\}$ is weakly stationary if:

(a) The mean is constant:

$$E[X_t] = \mu \quad \text{for all } t.$$

(b) The variance is constant:

$$\text{Var}(X_t) = \gamma(0) \quad \text{for all } t.$$

(c) The autocovariance depends only on the lag k (and not on time t):

$$\text{Cov}(X_t, X_{t+k}) = \gamma(k),$$

for all t and any lag k.

5. Let $\{X_t : t = 0, \pm 1, \pm 2, \ldots\}$ be a stochastic process given by;

$$X_t = e_t + c * t^2 + b + a$$

(for $t \geq 1$), where e_t is a white noise. Make X_t stationary by the application of the differencing approach (2 Marks)

Given the stochastic process:

$$X_t = e_t + c \cdot t^2 + b + a \quad (\text{for } t \geq 1)$$

where e_t is white noise.

The term $c \cdot t^2$ introduces a quadratic trend, which makes X_t non-stationary because the mean of X_t changes over time. To remove the trend and make X_t stationary, we can apply differencing. Differencing removes trends by subtracting the previous observation from the current observation.

First Differencing:

$$\Delta X_t = X_t - X_{t-1}$$

Substituting X_t:

$$\Delta X_t = (e_t + c \cdot t^2 + b + a) - (e_{t-1} + c \cdot (t-1)^2 + b + a)$$

Simplifying:

$$\Delta X_t = e_t - e_{t-1} + c \cdot (t^2 - (t-1)^2)$$

$$\Delta X_t = e_t - e_{t-1} + c \cdot (2t - 1) = e_t - e_{t-1} + 2ct - 2c$$

$$\Delta X_t = e_t - e_{t-1} + 2ct - c$$

Second Differencing:

The first differencing still leaves a linear trend $(2t - 1)$. To remove this, apply second differencing:

$$\Delta^2 X_t = \Delta X_t - \Delta X_{t-1}$$

Substituting ΔX_t:

$$\Delta^2 X_t = (e_t - e_{t-1} + c \cdot (2t - 1)) - (e_{t-1} - e_{t-2} + c \cdot (2(t-1) - 1))$$

Simplifying:

$$\Delta^2 X_t = e_t - 2e_{t-1} + e_{t-2} + 2c$$

After second differencing, the quadratic trend is removed, and $\Delta^2 X_t$ depends only on the white noise terms e_t, e_{t-1}, e_{t-2}, and a constant $2c$. Thus, $\Delta^2 X_t$ is stationary.

6. Under what situation is a time series referred to as a random walk? (1 Mark)
A time series $\{X_t\}$ is referred to as a **random walk** if it satisfies the following conditions:

$$X_t = X_{t-1} + e_t$$

where,e_t is a white noise process (i.e., e_t has zero mean, constant variance, and is uncorrelated over time) and The value of X_t at time t is equal to its value at time $t - 1$ plus a random shock e_t.

7. The random walk process is said to be non-stationary. Show why the random walk process described in (6) above is deemed a non-stationary time series process. (2 Marks)

Let $X_0 = 0$, $t = 0$

At $t = 1$, $X_1 = \epsilon_1$
$t = 2$, $X_2 = X_1 + \epsilon_2 = \epsilon_1 + \epsilon_2$

At $t = 3$, $X_3 = X_2 + \epsilon_3 = \epsilon_1 + \epsilon_2 + \epsilon_3$

At $t = 4$, $X_4 = X_3 + \epsilon_4 = \epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4$

By repeated substitution

$$X_t = \epsilon_1 + \epsilon_2 + \epsilon_3 + \ldots + \epsilon_t + X_0 = \sum_{t=1}^{t} \epsilon_t + X_0$$

Thus, effects of post errors remain in X_t. Therefore,

(a) $E(X_t) = \sum_{t=1}^{t} E(\epsilon_i)$=0

(b) $Var(X_t) = \sum_{i=1}^{t} Var(\epsilon_t) = \sum_{i=1}^{t} \sigma^2 = t\sigma^2$

Thus, a random walk is not stationary because it depends on time. However, differentiating moves a TS stationary.

8. Table 2 shows the average yearly production (in 10,000) of rice in the period 1995-2004. Using the data, calculate the following:

Table 2: Average monthly production

years (t)	1	2	3	4	5	6	7	8	9	10
X_t	70	76	76	70	53	52	42	56	63	66

(a) Expected function of X_t (2 Marks)

The expected function (mean) of X_t is calculated as:

$$\mathbb{E}[X_t] = \frac{1}{n} \sum_{t=1}^{n} X_t$$

$$\mathbb{E}[X_t] = \frac{70 + 76 + 76 + 70 + 53 + 52 + 42 + 56 + 63 + 66}{10} = \frac{624}{10} = 62.4$$

(b) Variance function of X_t (2 Marks)

The variance function is calculated as:

$$\text{Var}(X_t) = \frac{1}{n} \sum_{t=1}^{n} (X_t - \mathbb{E}[X_t])^2$$

Compute $(X_t - \mathbb{E}[X_t])^2$ for each t:

$$\begin{aligned}
(70 - 62.4)^2 &= 57.76 \\
(76 - 62.4)^2 &= 184.96 \\
(76 - 62.4)^2 &= 184.96 \\
(70 - 62.4)^2 &= 57.76 \\
(53 - 62.4)^2 &= 88.36 \\
(52 - 62.4)^2 &= 108.16 \\
(42 - 62.4)^2 &= 416.16 \\
(56 - 62.4)^2 &= 40.96 \\
(63 - 62.4)^2 &= 0.36 \\
(66 - 62.4)^2 &= 12.96
\end{aligned}$$

Sum of squared differences:

$$57.76+184.96+184.96+57.76+88.36+108.16+416.16+40.96+0.36+12.96 = 1153.4$$

$$\text{Var}(X_t) = \frac{1153.4}{10} = 115.34$$

(c) Autocovariance function at lag 1 (3 Marks)

The autocovariance function at lag k is given by:

$$\gamma(h) = \frac{1}{n}\sum_{t=1}^{n-k}(X_t - \mathbb{E}[X_t])(X_{t+k} - \mathbb{E}[X_t])$$

For lag $k = 1$:

$$\gamma(1) = \frac{1}{9}\sum_{t=1}^{9}(X_t - 62.4)(X_{t+1} - 62.4)$$

Compute each term:

$$\begin{aligned}
&(70 - 62.4)(76 - 62.4) = 7.6 \times 13.6 = 103.36\\
&(76 - 62.4)(76 - 62.4) = 13.6 \times 13.6 = 184.96\\
&(76 - 62.4)(70 - 62.4) = 13.6 \times 7.6 = 103.36\\
&(70 - 62.4)(53 - 62.4) = 7.6 \times (-9.4) = -71.44\\
&(53 - 62.4)(52 - 62.4) = (-9.4) \times (-10.4) = 97.76\\
&(52 - 62.4)(42 - 62.4) = (-10.4) \times (-20.4) = 212.16\\
&(42 - 62.4)(56 - 62.4) = (-20.4) \times (-6.4) = 130.56\\
&(56 - 62.4)(63 - 62.4) = (-6.4) \times 0.6 = -3.84\\
&(63 - 62.4)(66 - 62.4) = 0.6 \times 3.6 = 2.16
\end{aligned}$$

Sum of products:

$$103.36 + 184.96 + 103.36 - 71.44 + 97.76 + 212.16 + 130.56 - 3.84 + 2.16 = 759.04$$

$$\gamma_{(1)} = \frac{759.04}{9} = 84.34$$

(d) Autocovariance function at lag 2 (3 Marks)

For lag $k = 2$:

$$\gamma(2) = \frac{1}{8}\sum_{t=1}^{8}(X_t - 62.4)(X_{t+2} - 62.4)$$

Compute each term:

$$\begin{aligned}
&(70 - 62.4)(76 - 62.4) = 7.6 \times 13.6 = 103.36\\
&(76 - 62.4)(70 - 62.4) = 13.6 \times 7.6 = 103.36\\
&(76 - 62.4)(53 - 62.4) = 13.6 \times (-9.4) = -127.84\\
&(70 - 62.4)(52 - 62.4) = 7.6 \times (-10.4) = -79.04\\
&(53 - 62.4)(42 - 62.4) = (-9.4) \times (-20.4) = 191.76\\
&(52 - 62.4)(56 - 62.4) = (-10.4) \times (-6.4) = 66.56\\
&(42 - 62.4)(63 - 62.4) = (-20.4) \times 0.6 = -12.24\\
&(56 - 62.4)(66 - 62.4) = (-6.4) \times 3.6 = -23.04
\end{aligned}$$

Sum of products:

$$103.36 + 103.36 - 127.84 - 79.04 + 191.76 + 66.56 - 12.24 - 23.04 = 222.88$$

$$\gamma_{(2)} = \frac{222.88}{8} = 27.86$$

(e) Autocorrelation function at lag 1 (1 Mark)

The autocorrelation function at lag h is given by:

$$\rho_{(k)} = \frac{\gamma_{(k)}}{\gamma_{(0)}}$$

For lag $k = 1$:

$$\rho_{(1)} = \frac{\gamma_{(1)}}{\gamma_{(0)}} = \frac{84.34}{115.34} \approx 0.731$$

(f) Autocorrelation function at lag 2 (1 Mark)

For lag $k = 2$:

$$\rho_{(2)} = \frac{\gamma_{(2)}}{\gamma_{(0)}} = \frac{27.86}{115.34} \approx 0.242$$

@*TheAsamoah*