Statistics & Optimisation

Nishant Nayak
Msc. in Data Analytics
National College of Ireland
Dublin, Ireland
x22248242@student.ncirl.ie

***Abstract*— In part A of this study, we will show that Multiple Linear Regression is a statistical technique that is used to model the relationship between the dependent variable and the two or more independent variables. The primary approach of MLR is to fit a linear equation to observed data in such a way that it explains the variance in the dependent variable that can be associated with some predictor variables. This is often used in economics, biology and social sciences to understand and predict complex relationships.**

In part B of this study, we will perform Time Series Analysis. We must keep in mind that, the data points collected over time may have an internal structure (e.g., autocorrelation, trend, or seasonal variation) which ought to be accounted for.

Keywords— Multiple Linear Regression, Time Series Analysis

I. PART A- MULTIPLE LINEAR REGRESSION

The Multiple Linear Regression (MLR) is a statistical technique used for understanding the relationship between a dependent variable (also called an outcome or target variable) and multiple independent variables (also called predictors). This is done through a linear equation, where multiple predictors are used to explain the variance in the dependent variable; coefficients of the predictors are estimated to do this.

II. EXPLORATORY DATA ANALYSIS

A. Dataset Overview

Dataset Info: In the dataset, the numerical and categorical variables are (x1, x2, x3, y) → x3 is a categorical variable that needs to be encoded. Dataset Preview: Use info() and head() to inspect the data types and a sample of the data.

B. Missing Values

We need to do a quick check for any missing values using isnull().sum() and handle some columns by filling with column mean.

C. Feature Encoding

One-hot encode the categorical variable x3 into dummy variables (i.e. pd. (get_dummies) and eliminate the first column to prevent multicollinearity.

D. Numeric Conversion

Make sure all columns are numeric and handle any non-numeric data using pd.to_numeric(errors='coerce'). Remove rows with errors or NaN in them after conversion.

III. DATA PREPARATION

Before applying multiple linear regression, we need to prepare the data.

Cleaning the Data: Remove or impute missing values to ensure the dataset is complete.

Encoding Categorical Variables: If any categorical variables are present, convert them into numerical form. In this example, our dataset is purely numerical.

Feature Scaling: Standardise or normalise the features if they vary significantly in scale. This step ensures that all features contribute equally to the model.

A. Defining Variables

Independent Variables: X = [x1, x2, OneHot(x3_B), OneHot(x3_C)]

Dependent Variable (y): The target variable to be predicted.

B. Intercept Addition

Use sm to add a constant to the model for the intercept (sm.add_constant()).

C. Data Splitting

Using the train_test_split function, we will split the data into an 80:20 ratio for the training and test sets ensuring the split is reproduced by setting a random seed.

IV. MODELING AND INTERPRETATION

A. Model Fitting:

Implement MLR model on train using **sm.OLS(y_train, X_train).fit().**

Model Summary: Coefficients of predictors. p-values for statistical significance. Goodness of fit: adjusted R^2

B. Interpreting Results

P-values < 0.05 indicate significant predictors. The coefficient magnitude represents the effect of each predictor on y. The Adjusted R^2 measure indicates the amount of variance in y explained by the predictors.

V. DIAGNOSTICS AND EVALUATION

A. Residual Analysis

a) Normality of Residuals: : KDE with histogram of residuals.

b) Q-Q plot to ensure residual are normally distributed.

c) Homoscedasticity: Residuals vs Fitted Residuals need to disperse equally around zero.

B. Model Evaluation on Test Data

a) Make predictions for y for the test set.

b) Evaluate performance using: Mean Squared Error (MSE) which shows the average of the squares of the errors. R^2 Score: Explain the percentage of variance in y accounted for by the predictors.

C. Normality Test

a) Perform Shapiro-Wilk test for normality of the residuals. A p-value > 0.05 vanishes the residuals are well approximately normally distributed.

Below is the OLS Regression Results:

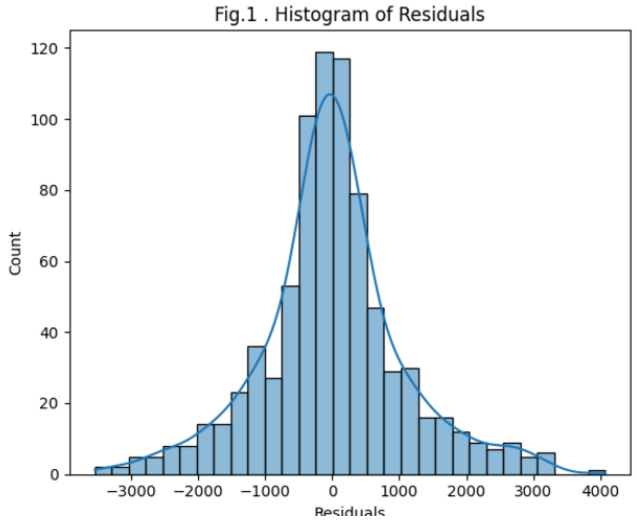
OLS Regression Results

Dep. Variable:	y	R-squared:	0.749
Model:	OLS	Adj. R-squared:	0.748
Method:	Least Squares	F-statistic:	593.4
Date:	Sun, 01 Dec 2024	Prob (F-statistic):	5.88e-237
Time:	00:08:39	Log-Likelihood:	-6717.9
No. Observations:	800	AIC:	1.345e+04
Df Residuals:	795	BIC:	1.347e+04
Df Model:	4		
Covariance Type:	nonrobust		

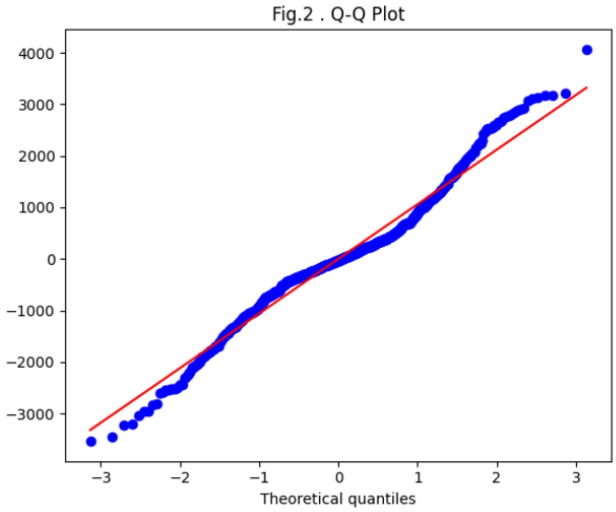
	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.539e+04	1753.721	-8.778	0.000	-1.88e+04	-1.2e+04
x1	301.9702	6.411	47.099	0.000	289.385	314.555
x2	79.3339	8.745	9.071	0.000	62.167	96.501
x3_B	-170.2221	91.539	-1.860	0.063	-349.908	9.464
x3_C	-91.6313	99.965	-0.917	0.360	-287.858	104.595

Omnibus:	27.400	Durbin-Watson:	2.024
Prob(Omnibus):	0.000	Jarque-Bera (JB):	58.230
Skew:	0.165	Prob(JB):	2.27e-13
Kurtosis:	4.280	Cond. No.	9.50e+03

b) Notes: Standard Errors assume that the covariance matrix of the errors is correctly specified. The condition number is large, 9.5e+03. This might indicate that there are strong multicollinearity.



Histogram of the residuals (Figure 1) suggests that they are close to being normally distributed but there are more residuals close to zero than perhaps you would expect.



Q-Q (quantile-quantile) plot assesses the normality of residuals from the multiple linear regression.

X-axis (Theoretical Quantiles): Represents the quantiles that are expected if the residuals are normally distributed.
Y-axis (Real quantiles): The quantiles of residuals.
Red Line: The best fit line for perfect normal distribution.
Blue Points: The residuals predicted.

Interpretation of the Plot:
Tight Fit to the Line: Residuals closely follow the normal distribution indicated at points along the red line. Deviations: Tails (Far Left and Right): magnitudes at both ends deviate a lot from the set of blue points. It sounds like there are some outliers or heavier tails in the residuals (which could indicate non-normality of the data). For near the center (middle quantiles), we observe that most points are close to red line, which means that the residuals are quite close to normal.

Conclusion: The residuals show deviations from normality, especially in the tails, which might signal problems such as: Outliers affecting the model. Model not capturing any skewed or any non-linear relationships.

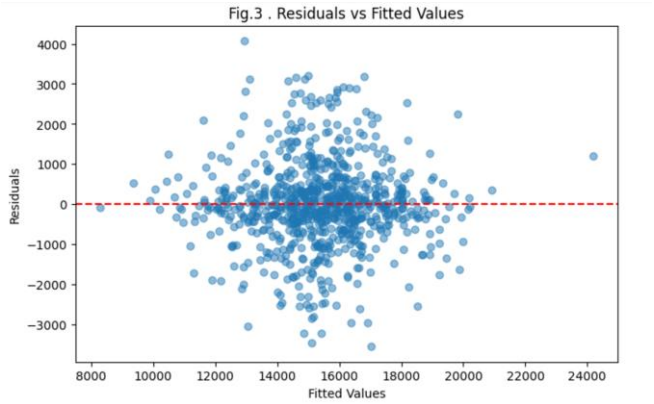


Fig.3 Represents a diagnostic plot that is used to check the assumptions of linear regression model.

The Residuals vs. Fitted Values plot is a diagnostic plot used to check the assumptions of linear regression. Below is a bit of the plot and what it means to the ML model:
Key Components of the Plot: Residuals –

Residuals (ei) are the differences between the observed (yi) and predicted (y^i) values

ei=yi−y^i

They represent the error in the model's predictions.

Fitted Values: Fitted values (y^i) are the predictions done by the regression model for the dependent variable.

Red Horizontal Line: The horizontal red line depicts ei=0, indicating perfect predictions with no residuals.

Insights from the Plot: Homogeneity of variance (homoscedasticity):

Observation about the plot: The residuals appear to be evenly distributed around the range of fitted values.

Constant variance of errors (If the variance increased or decreased in a systematic way (e.g., blended a shape of a funnel), it would be heteroscedastic.)

Centering Around Zero (Observation) : The residuals are centered around the red line (ei = 0).

The mean of the residuals is close to zero, and the model does not systematically overpredict or underpredict the dependent variable.

Random Scatter (Observation) : The points are randomly scattered, no pattern is visible.

Implication: This randomness means that the model is accurately describing the relationship between the independent variables and the dependent variable. The presence of any systematic pattern (e.g. curvature) would imply the model may lack important predictors or non-linear terms.

Outliers (Observation) : A few points are more distant from the ei=0 line.

Potential outliers mean that these values are predicted by the model but deviate significantly from the observed value. Outliers are expected within a dataset, but excessive outliers can damage the model and impact its performance

VI. Model Evaluation And Diagnostics Implications

A. *Linear Relationship*

The absence of any pattern in the residuals suggests that the relationship between dependent and independent variables is likely linear, reinforcing the case for a linear regression model. Residuals centered around zero and randomly scattered indicates that the model provides a reasonably good fit for the data.

B. *Areas to Investigate*

Just the few outliers need to be studied to see if they are true data or errors. If it were to show patterns or non-randomness, it would suggest that there was a need to:

1. Add more predictors
2. Transform variables
3. Consider alternative models – Non-Linear Regression or Machine Learning Algorithms.

VII. Mathematical Interpretation

A. *Homoscedasticity Assumption:*

$\mathbf{Var(ei)=\sigma 2\ \forall i}$

The variance of the residuals is the same for all fitted values.

B. *No Systematic Bias:*

$\mathbf{E(ei)=0\ \forall i}$
For instance, the mean of the residuals is zero, implying that predictions are not systematically biased.

C. *Randomness:*

Residuals (ei) should be uncorrelated with fitted values (y^i):
Cov (ei , y^i) = 0

VIII. Part A- Conclusion

The plot shows that the model fulfills the assumptions of a linear regression fairly well, especially homoscedasticity as well as there being no systematic bias. More advanced analysis may involve checking outliers further, and running statistical tests (e.g., Breusch-Pagan test for heteroscedasticity).

IX. Part B- Time Series Analysis

Time Series Analysis depends on the clue that the information focuses gathered throughout the years can have an internal structure (e.g., autocorrelation, trend or seasonal variation) that ought to be accounted for. Time series variables, for example exchange rates behave not it is irrational to forecast it and it is consistent. Despite these claims, most multinational firms, foreign-exchange dealers, exporters, importers and speculators still to decide on hedging based on expected rates with ex-post data as the basis of their decision. Such decisions are proposed with the assumption that there are patterns in the ex-post data and these patterns serve as an expected prediction for exchange rates behavior, at least in the short run. Therefore, if such patterns can be found, then in principle, they can be applied ARIMA, GARCH, and other advanced mathematical tools to learn and predict the ex-ante exchange rates (Hamilton 1994, Klaassen 1998).

The code forecasts a univariate time series using the ARIMA model. It follows these main steps:

1. Prepares and analyzes historical time series data.
2. Grid search of optimal ARIMA parameters
3. Estimate and evaluate an ARIMA model.
4. Predicts future values and displays the results.
5. It performs residual diagnostics on the model.

X. Exploratory Data Analysis

A. *Import necessary libaries:*

We make use of libraries like pandas, numpy, matplotlib, and statsmodels to manipulate, visualize and model the data.

B. *Load and preprocess the data:*

The dataset is read with pandas.read_csv() with Date as the index column for time series analysis.

C. *Cleaning the data:*

We will check and handle the missing values. Since there are no missing values in the dataset we will skip this step.

D. *Stationarity Check:*

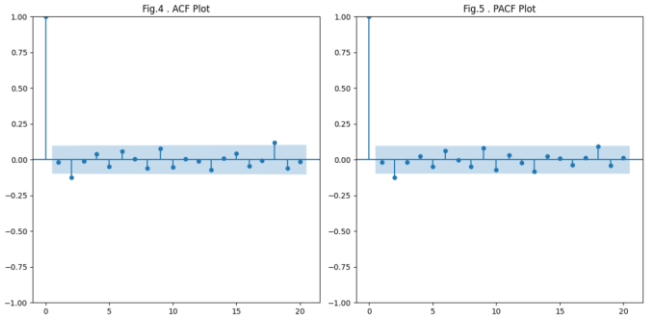
An Augmented Dickey-Fuller (ADF) test checks if the data is stationary. If not, we will need to transform the data using differencing.

E. *Differencing:*

To achieve stationarity, differencing is an essential step for ARIMA.
diff_data = data.diff().dropna()

F. *Plotting ACF and PACF:*

ACF method helps to understand the correlation between time series and its lagged versions (used to estimate q).
PACF helps to isolate individual lagged effects (used to estimate p).
plot_acf (diff_data , lags=20)
plot_pacf (diff_data , lags=20)



The diagram contains two sub-plots: an Autocorrelation Function (ACF) plot and a Partial Autocorrelation Function (PACF) plot. These are the tools commonly used in time series analysis for checking the autocorrelation of time series and selecting the appropriate lags for modeling.

1.ACF (Autocorrelation Function) — Left Plot
Definition: Computes the correlation of the time series with its lagged values at all time lags.

Features:
1.Lag 0 is equal to 1 because the series is perfectly correlated with itself.
2.The spikes correspond to autocorrelation values at different lags, while the shaded blue area gives the confidence interval (typically 95%).
3.Any peak above the shaded area is statistically significant.

Interpretation:

1.We see high autocorrelation at lag 0 via the plot
2.The lack of any spikes beyond lag 0 suggests that little or no autocorrelation exists beyond lag 0, with small insignificant spikes within the confidence interval.

2. PACF (Partial Autocorrelation Function) — Right Plot
Interpretation: This is a cross-correlation measure of how the time series correlates with its lagged values while controlling the contribution of intermediate lags.

Features:
1.Lag 0 also has a value of 1.
2.The spikes are partial autocorrelation at each lag.
3.The confidence interval is the shaded blue region, and large spikes are to fail out of it.
Interpretation:
1.The plot reveals a large peak at lag 0, with all other lags within the knife-edge confidence interval. This means that higher-order lags are not significantly contributing to the relationship.

XI. Data Preparation

A. *Grid search for ARIMA parameters:*

The function grid_search_arima() is used to iterate over possible values of p, d and q, fitting models and compare their AIC values (lower is better). The optimal parameters (best_order) are selected based on AIC.

B. *Parameter ranges:*

p_values = range (0,3)
d_values=range(0,1)
q_values = range (0,3)

XII. Modeling

A. *ARIMA Analysis*

ARIMA can be used to explain auto-correlations in the data. With ARIMA, stationary time series can be more precisely explained. Python is used for the implementation of monthly and annual time series in ARIMA.

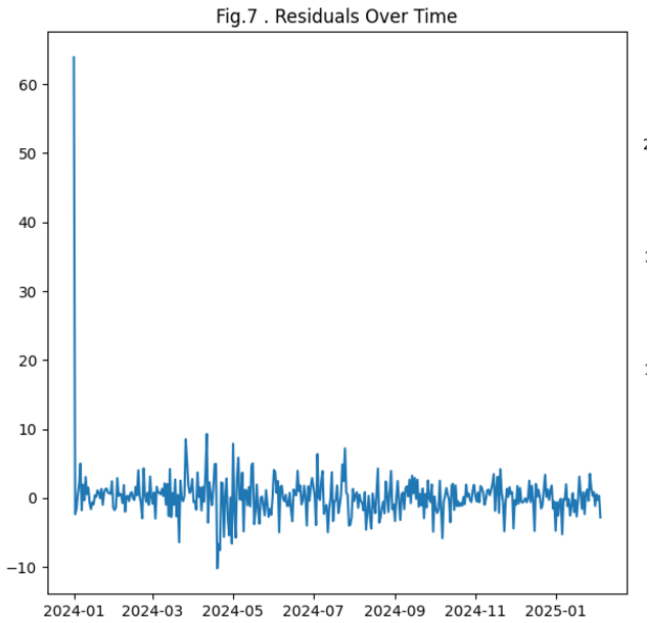


Fig.7 shows the residuals (differences between observed and predicted values) over time from an ARIMA model. Examining these residuals gives us valuable information on how well the ARIMA model fits the data and how well it conforms to its assumptions

Initial Spike:

At the beginning of the graph (January 2024), we have a huge spike in the residuals. This should be taken as a sign of instability in the initial predictions of the model, which could be caused by sudden changes in information or initialization problems.

Stabilization:

In the first period, we see an initial spike, but the residuals seem to stabilize, oscillating closer to zero, with a minor variance. This indicates that the model makes improvements in later stages.

Mean and Variance:

We observe the residuals to be centered around zero (mean $\approx$ 0), which is a desirable trait since it indicates no systematic prediction bias. Residuals variance appears stable over time (homoscedasticity), indicating the model's predictive utility.

Mean of residuals: 0.21241020613156755
Variance of residuals: 15.925873418530966

SARIMAX Results

Dep. Variable:	Value	No. Observations:	401
Model:	ARIMA(1, 1, 1)	Log Likelihood	-922.174
Date:	Mon, 02 Dec 2024	AIC	1850.347
Time:	19:22:58	BIC	1862.322
Sample:	01-01-2024 - 02-04-2025	HQIC	1855.089
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.5619	0.443	1.268	0.205	-0.307	1.430
ma.L1	-0.6102	0.429	-1.421	0.155	-1.452	0.231
sigma2	5.8883	0.304	19.369	0.000	5.292	6.484

Ljung-Box (L1) (Q):	0.25	Jarque-Bera (JB):	65.87
Prob(Q):	0.62	Prob(JB):	0.00
Heteroskedasticity (H):	0.40	Skew:	-0.12
Prob(H) (two-sided):	0.00	Kurtosis:	4.97

General Model Information:

Dependent Variable: The variable to be modeled (called "Value" here).

Model: ARIMA (1,1,1) — the model includes one lag of autoregressive (AR) and one lag for moving average (MA) with one order of differencing (d=1).

Number of Observations: 401 — The number of observations, or the number of data points in the time series used for fitting.

Sample: 01-01-2024 to 02-04-2025 — the range of the dataset.

Log Likelihood: -922.174 — assesses the fit. Larger values (nearer to 0) suggest a better fit.

AIC (Akaike Information Criterion): 1850.347 — used for model selection. These are penalized fit terms, so lower values indicate a better fit but also account for model complexity.

BIC (Bayesian Information Criterion): 1862.322 — like AIC but punishes additional parameters harder.

HQIC: 1855.089 — another variation on AIC.

Coefficients and Diagnostics:

The table shows the coefficients corresponding to the ARIMA model components along with their significance:

ar. L1 (Autoregressive term): Coefficient = 0.5619, P>|z| = 0.205. The p-value shows this term is not significant (greater than 0.05).

ma. L1 (Moving Average term): Coefficient = -0.6102, P>|z| = 0.155. As well, this term is not statistically significant.

Variance of residuals (sigma2): Coefficient = 5.8883, p-value (very low = 0.000). This term is quite important: here, the residual variance is being well estimated.

Model Diagnostics

Ljung-Box (Q) Test:

Q = 0.25 with Prob(Q) = 0.62. This test looks for autocorrelation in the residuals. If p-value > 0.05, it shows no significant autocorrelation, so at this moment, the model fits well regarding this.

Jarque-Bera (JB) Test:

JB = 65.87 with Prob(JB) = 0.00. This test verifies whether the residuals are normally distributed. A p-value of 0.00 indicates the residuals are not normal.

Heteroskedasticity (H) Test:

H = 0.40 with Prob(H) = 0.00. So, we are facing heteroskedasticity (non-constant variance of the residuals), which can impair the reliability of our model.

Skew: -0.12 — Asymmetry in residuals. A value close to zero indicates little skewness.

Kurtosis: 4.97 — indicates the "tailedness" of residuals. A value over 3 indicates heavy tails.

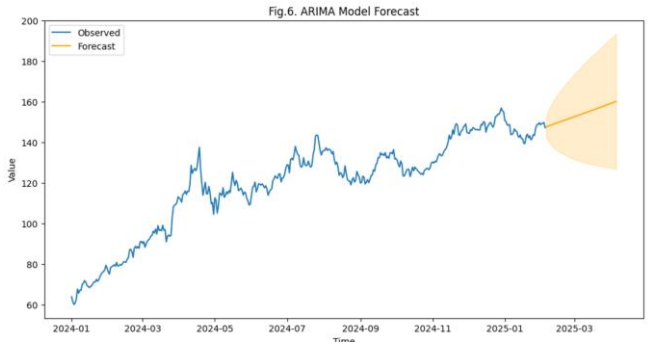


Fig.6 depicts an ARIMA model forecast for a time series.

Observed Data:

The blue line is the true historical data or actual values. Your time series starts in early 2024 and shows growth with some swings up and down over the year.

Forecast:

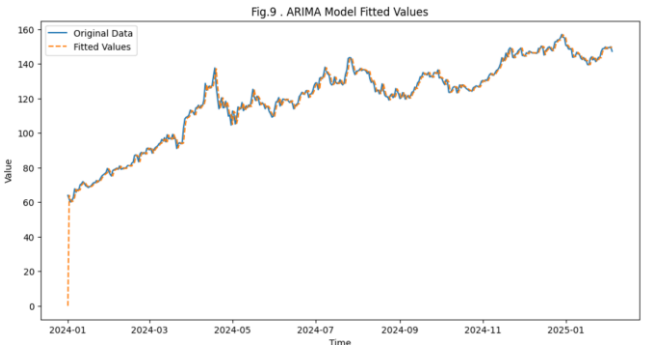
The orange line shows the predicted values coming from the ARIMA model, beginning just after the end of the observed data (around very late 2024). The trend has been upward, according to the model.

Uncertainty Region:

The darkened region around the forecast line is the confidence interval — how much the forecast spreads over time. This means that as the forecast period stretches further into the future (to March 2025), the predictions are becoming less certain.

Key Insights:

The projection shows consistent growth for the series value over time. The uncertainty cone visualizes the range of possibilities, providing a probabilistic look at the reliability of the forecast. To efficiently bring across the trends and uncertainties of the time series forecast, I have created this chart.



Fitted values of an ARIMA model over the time-series (against the time-series itself) Here's a rundown of some of the plot basics:

Blue Line (Original Data): These are the true values of the time series during the time period covered by the x-axis.
This is the benchmark with which the ARIMA model is compared.
Orange Dashed Line (Fitted Values): These are the forecasts made using the ARIMA model that we fit to the data.
The ARIMA model aims to reproduce the original data closely.
Interpretation: Here above, the orange dashed line follows the blue line very closely, which suggests that the ARIMA model is fitting the patterns and trends from the time series quite well.
However, there may be slight deviations, particularly when there are sudden peaks or troughs, since ARIMA models typically perform poorly in unstable or very turbulent data.
Time (X-Axis):

It denotes the temporal window for which data has been modeled.
But it is from the dates of January 2024 to early 2025 and shows how well the model fits all this time.
Value (Y-Axis): Represents the size of time series characteristic being acted upon.
For example, if the data is stock prices, then it will demonstrate the predicted value vs. the actual value of the stock.

Observations:

The ARIMA model appears to perform very well on trends and has significantly less error in fitted values, especially when trends are stable.
It may be slightly less accurate in cases of rapid increases or decreases (spikes or sharp dips).

B. SARIMA:

The SARIMA model has made an expansion to the SARIMA time series forecasting model. Forecasting using the SARIMA (auto-regressive integrated moving average) method results in projected values that are linear functions of the most recent actual values and recent prediction mistakes (residuals). Using the SARIMA model yields promising results, yet unsuitable for the next up time series. AIC, BIC, and HQIC are some informative metrics that are used for comparing diverse models. The lower the values of these criteria, the better the model is considered.

SARIMAX Results

Dep. Variable:	Value	No. Observations:	401
Model:	SARIMAX(1, 0, 2)	Log Likelihood	-914.950
Date:	Mon, 02 Dec 2024	AIC	1837.900
Time:	19:22:56	BIC	1853.846
Sample:	01-01-2024 - 02-04-2025	HQIC	1844.216
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.0014	0.001	1159.657	0.000	1.000	1.003
ma.L1	-0.0208	0.036	-0.572	0.568	-0.092	0.051
ma.L2	-0.1114	0.041	-2.739	0.006	-0.191	-0.032
sigma2	5.8115	0.298	19.531	0.000	5.228	6.395

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	61.96
Prob(Q):	0.99	Prob(JB):	0.00
Heteroskedasticity (H):	0.40	Skew:	-0.11
Prob(H) (two-sided):	0.00	Kurtosis:	4.92

The above image attached shows the SARIMAX (Seasonal ARIMA with exogenous regressors) model summary, which

gives statistical metrics along with estimated coefficients of each model component.
Summary of the model and key components:
Dependent Variable (Dep. Variable: Value): Shows the variable to be modeled or predicted.
Model Information: (Model: SARIMAX (1, 0, 2)):
p = 1: The notation ARI(p, d, q) describes an ARIMA model with p autoregressive (AR) terms.
d = 0: Indicates no differencing as the data is stationary.
q=2: Two moving average (MA) terms.

Observations:

The model was fitted to 401 observations.
Log Likelihood:

A statistic indicating how well the model fits data, where higher values mean better fit.
Akaike information criterion (aic): Information Criteria (AIC, BIC, HQIC)

Comparative Models—AIC (1837.900): Akaike Information Criterion (lower is better).
BIC (1853.846): Bayesian Information Criterion, imposes a larger penalty for model complexity than AIC.
HQIC (1844.216): Hannan-Quinn Information Criterion, a similar model selection metric.
Parameter Estimates
AR (Autoregressive Term):
ar.L1: Coefficient: 1.0014, very close to 1, which means that the time series is very persistent.
p-value: 0.000, indicating that the AR (1) term is statistically significant.
MA (Moving Average Terms):
ma.L1 (Lag 1): Coefficient: -0.0208, near 0, meaning this lag has little influence.
p-value: 0.568, nonsignificant.
ma.L2 (Lag 2):
Coefficient: -0.1114 => moderate 2nd lag impact
p-value = 0.006 (statistically significant).
Sigma2 (The variance of the residuals): Residual variance is estimated as 5.8115, which means the model is relatively less noisy.
Diagnostic Metrics:
Ljung-Box Test (L1: Q):

Q = 0.00, Prob(Q) = 0.99:
Suggests no significant autocorrelation in the residuals, hence a good model fit.
Jarque-Bera Test (JB):

JB = 61.96, Prob(JB) = 0.00:
Normality of residuals test. A small p
A non-normality if p-value < 0.05.
Heteroskedasticity (H):

H = 0.40, $Prob(H)$ = 0.00
H=0.40, Prob(H)=0.00:
Indicates residual heteroskedasticity.
⇒ Residual Distribution (Skew and Kurtosis):

Skew = -0.11: Residuals are slightly left-skewed.
Kurtosis = 4.92: The residuals have heavier tails than those for a normal distribution.
Key Observations

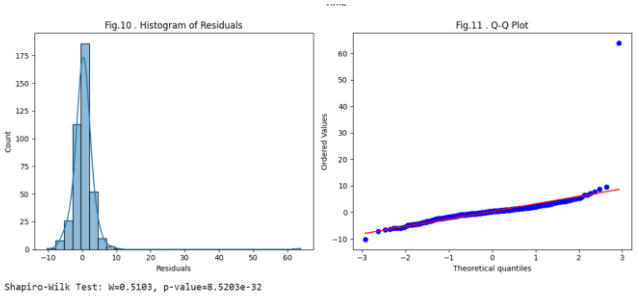
AR and MA Terms:

The AR (1) term is highly significant, and the MA(2) term is the only term that provides any contribution to the fit.

Good Model Fit:

If the residuals have low autocorrelation and the log-likelihood is reasonable, it's a good fit.
Issues:

Residuals are not completely normal (Jarque-Bera test fails) and there is some heteroskedasticity.



Two diagnostic plots are available in the image: Histogram of Residuals and Q-Q Plot. These are used to check the distribution of residuals (the error in the SARIMAX model). Below is an explanation for each plot:

Left Plot: Residuals Histogram

1. The histogram shows how the residuals (the errors of the SARIMAX model) are distributed.
2. A fitted density curve is overlaid to demonstrate the expected shape of the residuals distribution.

Key Observations:

1. Having the residuals clustered around 0 is what we want in a model that fits well.
2. But it is somewhat skewed and has a tail on the right side.
3. The residuals are also not exactly normally distributed.

Shapiro-Wilk Test:

1. The test statistic of the Shapiro-Wilk test (W = 0.5103) and p-value (8.5203×10^{-32}) are printed below the histogram.
2. A low p-value shows that the residuals are not normally distributed.

Right Plot: Q-Q Plot

1. A Quantile-Quantile (Q-Q) plot compares the quantiles of the residuals against the quantiles of a standard normal distribution.
2. The results superimposed show an ideal case where residuals correspond to a normal distribution, which is the red line.

Key Observations:

1. The points scatter closely around the red line, confirming information of approximate normality of the middle part of the residual distribution.
2. But at both extremes there are deviations:

- **Left tail:** There are points below the red line, which means there are heavier tails (outliers) on the negative side.
- **Right tail:** Points to the left of the line indicate larger positive outliers.

XIII. Part B: Conclusion

Non-Normal Residuals:

- Both the Shapiro-Wilk test and Q-Q plot confirm that we have deviance from normality in the residuals.
- This might affect statistical inferences like confidence intervals or hypothesis tests.

Mean Squared Error (MSE): 5.8029
Root Mean Squared Error (RMSE): 2.4089
These metrics are generally used to analyze time series models, RF, ARIMA, etc.
Mean Squared Error (MSE) is a value for measuring the average squared difference between observed actual outcomes and the outcomes predicted by the model. The lower the MSE, the better the performance of the model.

RMSE (Root Mean Squared Error) is just the square root of MSE, giving the error metric in the same units as the original data and thus making it easier to read.

In this case:
The root mean square error (RMSE) is an absolute measure of fit and should be in the same units as the outcome variable; it can represent how much the clustered model needs to be shifted in order to obtain the predicted value on average, indicating that the predicted values are an average of 2.4089 units away from the true values.

XIV. References

[1] Southampton, U. (2015) *Module 3 - Multiple Linear Regressions - Restore*, *Restore National Centre*. Available at:https://www.restore.ac.uk/srme/www/fac/soc/wie/research-new/srme/modules/mod3/module_3_ _multiple_linear_regression.pdf (Accessed: 01 December 2024).

[2] Djanie, T. and Thompson, H.D. (2012) *STAT 758 Time Series Project*. Available at: https://www.cse.unr.edu/~harryt/CS773C/Project/Time%20series%20project.pdf (Accessed: 01 December 2024).