ASHISHKUMAR PEMMARAJU

DS 480 FINAL PROJECT REPORT

WIU ID: 923-19-4632

The data utilized originates from two different R libraries, Time Series Analysis (TSA) and Applied Statistical Time Series Analysis (ASTSA). TSA includes the Monthly Electricity Production Time Series data, employed for ARIMA and SARIMA modeling. Conversely, ASTSA incorporates datasets such as the Monthly Export Price of Salmon and Returns on the New York Stock Exchange, which are utilized for GARCH and ARCH modeling, respectively. Platelet Levels data is used for State Space Models.

ARIMA MODELLING

The Autoregressive Integrated Moving Average (ARIMA) model is a popular time series forecasting technique that combines autoregression, differencing, and moving averages. Here's a breakdown of its description, limitations, uses, and examples:

1. **Autoregression** (AR): The AR part of ARIMA indicates that the evolving variable of interest is regressed on its own lagged values. It incorporates the linear relationship between an observation and a number of lagged observations (also known as its "lags").

2. **Integrated** (I): The I part of ARIMA indicates that the data values have been replaced with the difference between their values and the previous values (i.e., differencing). Differencing helps to make the time series stationary, which is often necessary for modeling.

3. **Moving Average** (MA): The MA part of ARIMA indicates that the regression error is a linear combination of error terms whose values occurred contemporaneously and at various times in the past.

The ARIMA model is typically denoted as **ARIMA (p, d, q)**, where:

- p is the order of the autoregressive part,

- d is the degree of differencing, and

- q is the order of the moving average part.

> Limitations:

1. **Assumption of Linearity**: ARIMA assumes that the relationship between the variable of interest and its past values is linear. If the relationship is nonlinear, ARIMA might not perform well.

2. **Stationarity Requirement**: ARIMA requires the time series to be stationary or made stationary through differencing. Some time series data may be difficult to stationarize, which can limit the model's effectiveness.

3. **Lack of Seasonality Handling**: Standard ARIMA models do not directly handle seasonal variations in the data. Seasonal ARIMA (SARIMA) models are often used to address this limitation.

> Uses:

1. **Time Series Forecasting**: ARIMA models are widely used for forecasting future values of a time series variable based on its historical values.

2. **Financial Forecasting:** ARIMA models are applied in financial markets for predicting stock prices, exchange rates, and other financial variables.

3**. Economic Analysis**: ARIMA models are used in economic analysis for forecasting macroeconomic indicators such as GDP, inflation, and unemployment rates.

4. **Demand Forecasting**: ARIMA models are employed in industries such as retail and manufacturing for forecasting product demand.

> Examples:

1. Stock Price Prediction: Predicting future stock prices based on historical stock price data.

2. Sales Forecasting: Forecasting future sales of a product based on historical sales data.

3. Temperature Forecasting: Predicting future temperatures based on historical temperature data.

4. Traffic Forecasting: Forecasting future traffic volume on a particular road based on historical traffic data.

ARIMA models are versatile and widely used in various fields for time series forecasting tasks. However, their performance can be influenced by the quality and characteristics of the data, as well as the appropriateness of model selection.

- **MATHEMATICAL FORMULATION**

$y_t' = c + \phi_p \cdot y_{t-p} + \phi_{p-1} \cdot y_{t-(p-1)} + \cdots + \phi_1 \cdot y_{t-1} + \theta_q \cdot \varepsilon_{t-q} + \theta_{q-1} \cdot \varepsilon_{t-(q-1)} + \cdots + \theta_1 \cdot \varepsilon_{t-1} + \varepsilon_t$
Where: y_t' represents the differenced series

'p' is the autoregressive order.

C is the constant term.

'd' is the differencing order

Φ and θ represents Autoregressive and moving average coefficients respectively.;

'q' is the moving average order.

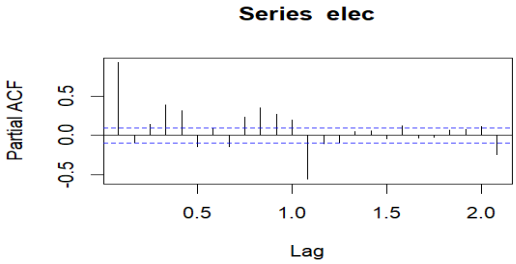
yt represents the time series data at Time.

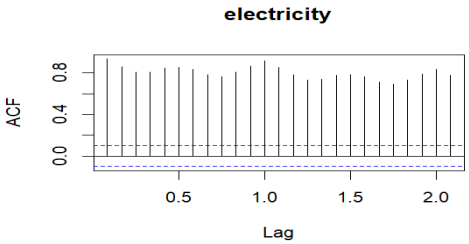
Et represents the error term/residuals at time t

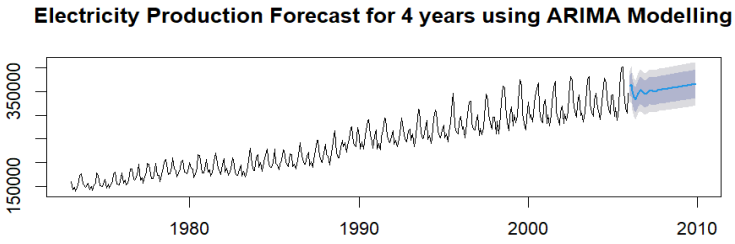
Monthly Electricity Production in USA

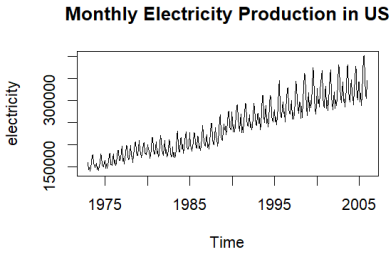
ARIMA Output

```
Series: elec
ARIMA(3,1,1) with drift

Coefficients:
         ar1      ar2      ar3      ma1    drift
      0.5850  -0.2201  -0.3136  -0.9362  484.952
s.e.  0.0494   0.0552   0.0492   0.0180   57.460

sigma^2 = 269458149:  log likelihood = -4393.29
AIC=8798.57   AICc=8798.79   BIC=8822.45

Training set error measures:
                    ME     RMSE      MAE        MPE     MAPE     MASE       ACF1
Training set -75.90402 16290.35 12154.93 -0.4940248 4.991111 1.397261 0.02436394
```









- ACF plot indicates all positive values indicating that there has always been electricity production every month without interruption.
- Monthly production plot indicates a regular increase in the production as the time periods increase indicating and linear and increasing trend.
- Forecast of the next 4 years or 48 months provides us with an idea of how the next years could be having more electricity production. The highlighted section of the forecast graph represents the forecasted values and trends in the upcoming 48 months, the plot displays a stable trend with time.
- In summary, the ARIMA (3,1,1) model has been configured with specific parameters indicating the number of autoregressive (AR) terms, differencing (I), and moving average (MA) terms. Estimated coefficients for AR and MA terms are provided, with positive and negative effects. Standard errors (S.E.) represent the precision of these estimates.
- The variance of residuals (sigma^2) shows unexplained variability, while log likelihood indicates model fit. Lower values of AIC, AICC, and BIC suggest better models, balancing goodness of fit and complexity. Error measures like ME, RMSE, MAE, MPE, and MASE assess model performance. ACF1 measures autocorrelation of residuals at lag1, with values close to 0 suggesting no significant autocorrelation between consecutive residuals.

SARIMA MODEL

Seasonal Autoregressive Integrated Moving Average (SARIMA) is an extension of the ARIMA model that incorporates seasonality. SARIMA models are designed to handle time series data with seasonal patterns. Here's an overview of SARIMA, including its description, limitations, uses, and examples:

Description:

1. **Seasonal Component**: SARIMA includes additional parameters to model seasonal variations in the data. It captures the seasonal patterns by incorporating seasonal autoregressive (SAR) and seasonal moving average (SMA) terms, in addition to the non-seasonal ARIMA components.

2. **Integration**: Like ARIMA, SARIMA includes differencing to achieve stationarity. The degree of differencing (d) in SARIMA represents both non-seasonal and seasonal differences, if necessary.

3. **Model Notation**: SARIMA models are typically denoted as SARIMA(p, d, q)(P, D, Q, s), where:

- (p, d, q) are the non-seasonal ARIMA parameters,
- (P, D, Q) are the seasonal ARIMA parameters,
- s is the length of the seasonal cycle (e.g., 12 for monthly data).

Limitations:

1. **Complexity**: SARIMA models can become complex, especially for datasets with multiple seasonal cycles. Estimating the parameters of such models may require large amounts of data and computational resources.

2. **Interpretability**: SARIMA models with many parameters can be difficult to interpret, making it challenging to understand the underlying patterns in the data.

Uses:

1. **Seasonal Forecasting**: SARIMA models are used for forecasting time series data that exhibit seasonal patterns, such as monthly or quarterly data.

2. **Retail Sales Forecasting**: SARIMA models are applied in retail industry for predicting seasonal fluctuations in sales of products.

3. **Energy Demand Forecasting**: SARIMA models are used by utilities to forecast energy demand, which often exhibits seasonal patterns.

Examples:

1. **Monthly Sales Prediction**: Forecasting monthly sales of a product based on historical sales data, considering seasonal variations due to holidays or other factors.

2. **Quarterly Revenue Forecasting**: Predicting quarterly revenue of a company based on historical financial data, accounting for seasonal fluctuations in sales.

3. **Weather Forecasting**: Forecasting seasonal weather patterns, such as temperature and precipitation, using historical weather data.

4. **Tourism Demand Forecasting**: Predicting seasonal variations in tourism demand for a particular destination based on historical visitation data.

SARIMA models are powerful tools for modeling and forecasting time series data with seasonal patterns. They allow analysts to capture and account for the cyclicality inherent in many real-world datasets. However, as with any modeling approach, careful consideration should be given to the data characteristics and model selection process to ensure reliable forecasts.

- **MATHEMATICAL FORMULATION**

$(1-\phi 1B-\phi 2B2-\cdots-\phi pBp) \times(1-\Phi 1Bs-\Phi 2B2s-\cdots-\Phi PBPs) \times(1-B)d\times(1-Bs)Dyt=(1+\theta 1B+\theta 2B2+\cdots+\theta qBq)\times(1+\Theta 1Bs+\Theta 2B2s+\cdots+\Theta QBQs)\times\varepsilon t$

> Where:

Yt is the observed time series.

B represents the operator that shifts the time series by one period.

$\phi 1,\phi 2$ represents the non-seasonal autoregressive parameters.

$\Phi 1,\Phi 2$ represents the seasonal autoregressive parameters.

$\theta 1,\theta 2$ represents the non-seasonal moving average parameters.

$\Theta 1,\Theta 2$ represents the seasonal moving average parameters.

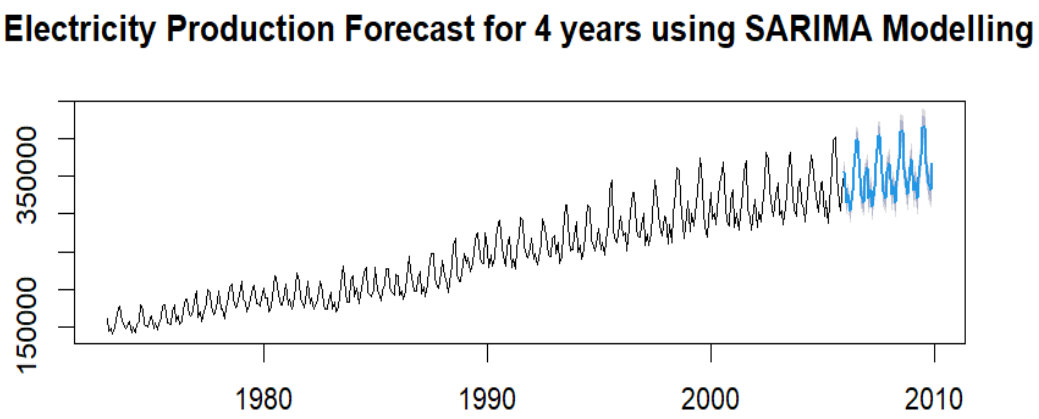
εt err refers to the error term.

- **SARIMA Modelling on Monthly Electricity Production Data Result**

```
Series: elec
ARIMA(1,0,2)(0,1,1)[12] with drift

Coefficients:
         ar1      ma1      ma2     sma1     drift
      0.9526  -0.4213  -0.2888  -0.6980  480.6844
s.e.  0.0248   0.0571   0.0564   0.0339   56.5038

sigma^2 = 49279084:  log likelihood = -3947.21
AIC=7906.42   AICc=7906.64   BIC=7930.12

Training set error measures:
                    ME     RMSE      MAE        MPE     MAPE      MASE         ACF1
Training set -20.11085 6867.574 5177.861 -0.1414256 2.085157 0.5952172 0.005558972
```

- The ARIMA model configuration includes autoregressive (AR) terms AR1 and AR2, along with seasonal autoregressive (SAR) terms SAR1 and SAR2, indicating a seasonal period of 12.
- Additionally, there's a seasonal moving average term SMA1 and a drift term for a constant linear trend. Standard errors (S.E.) reflect the precision of estimated coefficients.
- The variance of residuals (sigma^2) captures unexplained variability. Log likelihood assesses model fit, with higher values indicating better fit. AIC, AICC, and BIC gauge model complexity, favoring lower values.
- Error measures like ME, RMSE, MAE, MPE, and MASE evaluate model performance. ACF1 measures autocorrelation at lag1, with values near 0 suggesting negligible autocorrelation between consecutive residuals.

PLOTS



- Comparing the forecast plots of ARIMA and SARIMA, we can observe a slight difference in the predicted amount of production, indicating that the seasonal effect on the electricity production has a significant impact on the produce.

Reason To Use ARIMA And SARIMA For Predicting Electricity Production

- ARIMA and SARIMA models provide precise forecasts for monthly electricity production in the US by accounting for seasonal variations, trends, and short-term fluctuations.
- With their flexibility and interpretability, these models enable stakeholders to anticipate changes in energy demand, optimize resource allocation, and develop effective energy policies.
- Additionally, their historical performance in forecasting time series data underscores their reliability and utility in addressing the dynamic nature of electricity production.
- Thus, ARIMA and SARIMA models serve as invaluable tools for informed decision-making and strategic planning in the energy sector**.**

ARCH AND GARCH MODELS

ARCH (Autoregressive Conditional Heteroskedasticity) and GARCH (Generalized Autoregressive Conditional Heteroskedasticity) models are statistical models used in finance to model and forecast the volatility of financial time series data.

ARCH: ARCH models, introduced by Robert Engle in 1982, model the conditional variance of a time series based on past squared errors or residuals. They assume that volatility clusters in time, meaning periods of high volatility tend to be followed by periods of high volatility.

GARCH: GARCH models, developed by Tim Bollerslev in 1986, extend ARCH models by incorporating past conditional variances in addition to past squared errors. This allows GARCH models to capture both short-term and long-term volatility effects, making them more flexible and accurate for modeling financial volatility.

ARCH (Autoregressive Conditional Heteroskedasticity) and GARCH (Generalized Autoregressive Conditional Heteroskedasticity) models have various applications in finance and econometrics:

Uses:

1. **Volatility Forecasting**: Both ARCH and GARCH models are primarily used for forecasting the volatility of financial time series data. This is crucial for risk management, asset pricing, and portfolio optimization.

2. **Risk Management**: These models help financial institutions and investors understand and manage risk by providing estimates of future volatility. For example, in portfolio management, accurate volatility forecasts are essential for calculating Value at Risk (VaR) and Expected Shortfall (ES).

3. **Option Pricing**: Volatility is a key input in option pricing models like the Black-Scholes model. ARCH and GARCH models provide more accurate estimates of future volatility, leading to better pricing of options and other derivatives.

4. **Financial Market Analysis**: These models are used to analyze the behavior of financial markets and identify patterns in volatility. They help investors make informed decisions by providing insights into market dynamics.

Limitations:

1. **Normality Assumption**: They assume normal distribution, but financial data often deviates from this, impacting accuracy.

2. **Model Complexity**: Higher-order models require substantial data and computational resources, increasing the risk of overfitting.

3. **Model Instability**: They may not capture sudden changes in volatility during market shifts, affecting performance.

Examples:

1. Non-Normality in Financial Returns: Financial returns often exhibit non-normal distributions, such as fat tails and skewness. For instance, during market crashes or extreme events, returns may deviate significantly from a normal distribution, violating the assumption of normality in ARCH and GARCH models. This can lead to inaccurate volatility forecasts and risk estimates, potentially underestimating the true level of risk in financial markets.

2. Model Overfitting: When using higher-order ARCH and GARCH models with many parameters, there is a risk of overfitting the data. Overfitting occurs when the model captures noise or random fluctuations in the data, rather than underlying patterns or relationships. As a result, the model may perform well on the training data but poorly on new, unseen data. This can lead to misleading volatility forecasts and unreliable risk management decisions.

- **Mathematical Formulae:**

ARCH Model

$\sigma_t^2 = \alpha_0 + \sum^{p}_{i=1} \alpha_i \cdot \varepsilon^2_{t-i}$

where:

- σ_t^2 denotes the conditional variance at time t
- α_0 denotes a constant term representing intercept.
- α_i represents ARCH coefficients for lags 1 to P.
- ε^2_{t-i} denotes the squared residual errors at lag i.

GARCH Model

GARCH Model can be expressed mathematically as:

$\sigma_t^2 === \omega + \sum^{p}_{i=1} \alpha_i \cdot \varepsilon^2_{t-I} + \sum^{q}_{j=1} \beta_j \cdot \sigma^2_{t-j}$

where:

- ω is a constant term representing intercept of the variance equation.
- α_i and β_j represents ARCH and GARCH coefficients for lags 1 p and q respectively.
- ε^2_{t-I} denotes the squared residual errors at lag i.
- σ^2_{t-j} denotes the conditional variance at lag j.

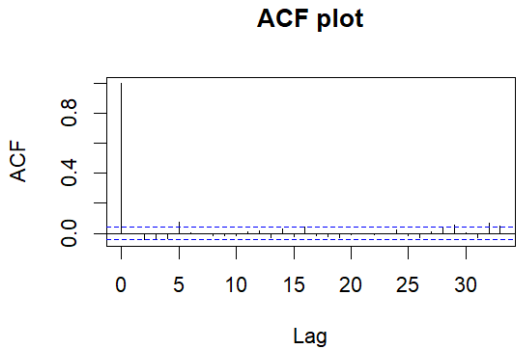
> ARCH AND GARCH MODELLING FOR RETURNS ON NEW YORK STOCK EXCHANGE OUTPUT:

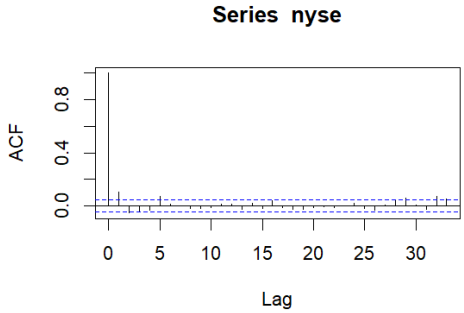
```
*----------------------------------*
*          GARCH Model Fit         *
*----------------------------------*

Conditional Variance Dynamics
-----------------------------------
GARCH Model  : sGARCH(3,4)
Mean Model   : ARFIMA(1,0,1)
Distribution : norm

Optimal Parameters
------------------------------------
        Estimate  Std. Error   t value Pr(>|t|)
mu      0.000678    0.000197  3.437672 0.000587
ar1    -0.095105    0.250134 -0.380216 0.703785
ma1     0.203571    0.244925  0.831159 0.405884
omega   0.000009    0.000000 68.465317 0.000000
alpha1  0.161490    0.025104  6.432947 0.000000
alpha2  0.000002    0.103730  0.000016 0.999987
alpha3  0.000000    0.099158  0.000003 0.999998
beta1   0.218849    0.590966  0.370323 0.711142
beta2   0.000005    0.504087  0.000009 0.999992
beta3   0.299942    0.140711  2.131618 0.033038
beta4   0.205037    0.155010  1.322730 0.185925

Robust Standard Errors:
        Estimate  Std. Error   t value Pr(>|t|)
mu      0.000678    0.000288  2.353046  0.01862
ar1    -0.095105    0.279609 -0.340135  0.73376
ma1     0.203571    0.283041  0.719228  0.47200
omega   0.000009    0.000000 25.072373  0.00000
alpha1  0.161490    0.113819  1.418828  0.15595
alpha2  0.000002    0.535071  0.000003  1.00000
alpha3  0.000000    0.468630  0.000001  1.00000
beta1   0.218849    3.048570  0.071787  0.94277
beta2   0.000005    2.331746  0.000002  1.00000
beta3   0.299942    0.341199  0.879082  0.37936
beta4   0.205037    0.479178  0.427893  0.66873

LogLikelihood : 6740.096

Information Criteria
------------------------------------

Akaike       -6.7291
Bayes        -6.6983
Shibata      -6.7292
Hannan-Quinn -6.7178

Weighted Ljung-Box Test on Standardized Residuals
------------------------------------
                        statistic p-value
Lag[1]                      1.420  0.2334
Lag[2*(p+q)+(p+q)-1][5]     2.329  0.8602
Lag[4*(p+q)+(p+q)-1][9]     3.082  0.8778
```

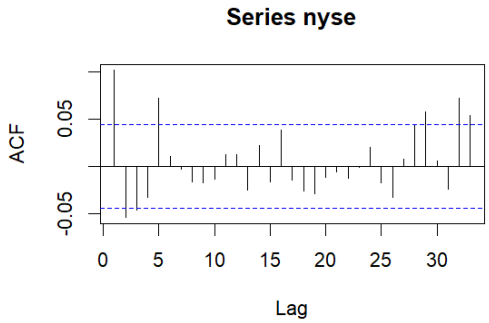
```
d.o.f=2
H0 : No serial correlation

Weighted Ljung-Box Test on Standardized Squared Residuals
------------------------------------
                         statistic p-value
Lag[1]                      0.7006  0.4026
Lag[2*(p+q)+(p+q)-1][20]    4.1466  0.9823
Lag[4*(p+q)+(p+q)-1][34]    7.6500  0.9941
d.o.f=7

Weighted ARCH LM Tests
------------------------------------
             Statistic Shape Scale P-Value
ARCH Lag[8]     0.1275 0.500 2.000  0.7210
ARCH Lag[10]    1.3649 1.488 1.815  0.6769
ARCH Lag[12]    1.8185 2.451 1.700  0.8199

Nyblom stability test
------------------------------------
Joint Statistic:  93.8943
Individual Statistics:
mu     0.1231
ar1    0.2035
ma1    0.2071
omega  2.4508
alpha1 0.1820
alpha2 0.2654
alpha3 0.2462
beta1  0.2255
beta2  0.2157
beta3  0.1977
beta4  0.2063

Asymptotic Critical Values (10% 5% 1%)
Joint Statistic:          2.49 2.75 3.27
Individual Statistic:     0.35 0.47 0.75

Sign Bias Test
------------------------------------
                   t-value    prob sig
Sign Bias          0.1426 0.88664
Negative Sign Bias  2.1953 0.02826  **
Positive Sign Bias  1.0995 0.27166
Joint Effect        9.0324 0.02886  **


Adjusted Pearson Goodness-of-Fit Test:
------------------------------------
  group statistic p-value(g-1)
1    20     95.96    2.867e-12
2    30    100.33    8.663e-10
3    40    123.96    8.835e-11
4    50    125.55    1.209e-08


Elapsed time : 0.514149
```

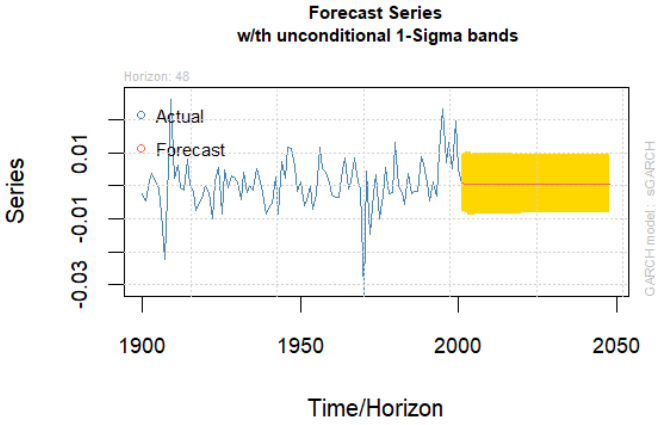
- **Output description:**
- The output represents the results of fitting a GARCH model to some financial or time series data. The model specifications include a symmetric GARCH (3,4) model with an ARIMA (1,0,1) mean model and a normal distribution assumption.
- The estimated parameters, along with their standard errors and significance levels, are provided for the mean, autoregressive, moving average, ARCH, and GARCH components of the model.
- The log-likelihood value indicates the goodness of fit, while various information criteria aid in model selection. Diagnostic tests assess the adequacy of the model, including tests for autocorrelation, ARCH effects, stability, asymmetry, and overall goodness-of-fit.
- This comprehensive summary allows for the evaluation of the model's performance and suitability for the given data.

- **Plots**









- The ACF plot of residuals examines the correlation between residuals at different time lags, revealing any remaining autocorrelation that could signal model inaccuracies.
- Values near zero signify uncorrelated residuals at a given lag. Notably, significant spikes at certain lags, indicate strong positive correlation, while others, suggest strong negative correlation.
- This plot aids in assessing whether the ARCH Model adequately captures temporal patterns in the data. Significant correlations suggest potential model shortcomings, while non-significant autocorrelation implies the model effectively captures temporal patterns.
- The forecast plot shows predicted values for the next 48 months, highlighting future trends. The stable trend over time is evident in the highlighted section of the plot, indicating consistent forecasts for the upcoming 48 months.
- **Purpose:**

ARCH and GARCH models are beneficial for predicting returns in the New York Stock Exchange because they capture volatility dynamics, account for conditional heteroskedasticity, improve risk management, inform trading strategies, and evaluate the risk-return tradeoff.

STATE SPACE MODEL

A state space model is a mathematical framework used to represent and analyze dynamic systems. It consists of two main components: the state equation and the observation equation.

1. **State Equation**: This equation describes the evolution of the underlying system over time. It represents how the system's internal state changes from one time step to the next. The state equation is typically formulated as a linear or nonlinear dynamic process, often in the form of a transition equation. It defines the relationship between the current state of the system and its previous state, as well as any external inputs or disturbances affecting the system dynamics.

2**. Observation Equation**: This equation relates the observable outputs of the system to its internal state. It describes how measurements or observations of the system are generated from its underlying state. The observation equation can be linear or nonlinear and may include additive noise or measurement error terms.

> Uses:

1. **Time Series Analysis**: For forecasting and smoothing time-varying processes like economic indicators and weather patterns.

2. **Control Systems**: In designing control algorithms for systems such as industrial processes and robotics.

3. **Signal Processing**: For tasks like speech recognition, image processing, and sensor data fusion.

4. **Econometrics and Finance**: For modeling economic and financial time series data, forecasting indicators, and analyzing markets.

> Limitations:

1. **Complexity**: State space models can become complex, especially for systems with many variables or nonlinear dynamics. This complexity can make model estimation and interpretation challenging, particularly when dealing with high-dimensional data or intricate system behaviors.

2. **Model Uncertainty:** State space models rely on assumptions about the underlying system dynamics and measurement processes. If these assumptions are incorrect or incomplete, the model may produce unreliable results. Uncertainty in model parameters and structures can lead to inaccurate predictions and unreliable inference.

3. **Computational Demands**: Estimating state space models often requires iterative numerical techniques such as the Kalman filter or particle filtering methods. These algorithms can be computationally demanding, especially for large datasets or complex models. Additionally, model selection and parameter estimation may require extensive computational resources and time, limiting the practical applicability of state space models in some scenarios.

> Mathematical Formulae:

The representation of a state space model involves two main equations, the observation and state equation.

STATE EQUATION

$$x_t = F_t x_{t-1} + G_t w_t$$

x_t denotes the state vector at time t.

F_t is the transition matrix that describes how the state evolves t-1 to t.

G_t is the matrix relating stochastic noise term w_t to state transition.

w_t is the noise assumed to be a zero mean random vector with a known covariance matrix Q_t.

OBSERVATION EQUATION

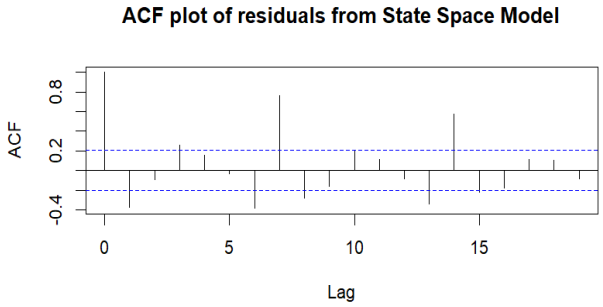
$y_t = H_t x_t + v_t$

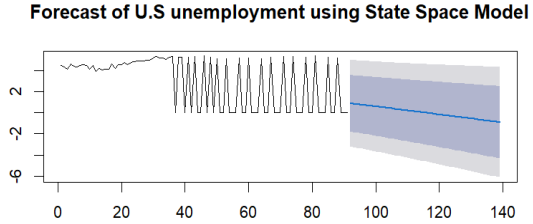
y_t denotes observed measurement at time t.

H_t represents the observation matrix that relates the state x_t to the observed measurements.

v_t denotes the measurement noise assumed to be zero-mean random vector with a known covariance matrix.

PLOTS:





Plot derivation:

- The ACF (Autocorrelation Function) plot of the residuals generated by this model allows us to visualize the autocorrelation structure within the residuals. It helps us assess whether there are any significant correlations between the residuals at different time lags.
- In this specific case, there was a notable spike at the first lag, indicating a strong correlation at that lag. This suggests that there may be some remaining patterns or dependencies in the data that the model has not fully captured.
- On the other hand, the forecast plot provides insights into the predicted trends for the next 48 months, equivalent to 4 years, based on the state space model. It visualizes the forecasted values along with the associated uncertainty, typically represented by prediction intervals.
- In this case, the forecasted values are derived from the state space model, incorporating both the model's estimated parameters and any uncertainty in those estimates.
- The plot typically includes low and high prediction intervals, often at 80% and 95% confidence levels, to illustrate the range of possible outcomes. This allows decision-makers to gauge the level of confidence in the forecast and assess potential risks associated with future outcomes.