Project Report on

TIME SERIES ANALYSIS

Submitted by:

Name: **Vaishnavi Pandey**
Integrated MSc. Statistics
Central University of Rajasthan

Submitted to:

Prof. Nirpeksh Kumar
Department of Statistics
Banaras Hindu University

Acknowledgement

I want to extend a sincere and heartfelt obligation towards all the personages without whom the completion of the internship project was not possible. I express my profound gratitude and deep regard to Prof. Nirpeksh Kumar , BHU for his guidance, valuable feedback and constant encouragement throughout the project. His valuable suggestions were of immense help. I sincerely acknowledge his constant support and guidance during the project.

I am also grateful to the Banaras Hindu University for allowing me to this project and providing all the required facilities.

Introduction

During my internship, I had the opportunity to conduct a comprehensive time series analysis on the number of births per month in New York City from January 1946 to December 1959. This problem aimed to uncover underlying trends, seasonal patterns, and make future forecasts using a variety of statistical techniques.

Time series analysis is a powerful tool for examining how data points are collected or recorded at specific time intervals. It enables us to identify patterns and relationships within the data, which can be used for forecasting and making informed decisions. In this analysis, I applied several techniques to decompose the time series into its components: trend, seasonality, and noise. Additionally, I explored methods for detrending, deseasonalizing, and stabilizing the variance of the series to improve the accuracy of our models.

The dataset for this analysis was sourced from Rob J Hyndman's time series data library. The series consists of monthly birth counts in New York City over a 14-year period. By analyzing this data, we aim to understand the historical patterns and predict future birth trends, which can be valuable for city planning, healthcare resource allocation, and other policy-making decisions.

This report documents the entire process of the analysis, from initial data visualization to advanced forecasting using ARIMA models. It includes detailed explanations of the statistical techniques used, the rationale behind each step, and the findings derived from the analysis. Through this problem, I enhanced my understanding of time series analysis and gained practical experience in applying statistical methods to real-world data.

Objective

The primary objective of this internship report is to document the comprehensive time series analysis conducted on the number of births per month in New York City from 1946 to 1959. This report aims to detail the methodologies, theoretical concepts, and findings from the analysis, demonstrating the application of statistical techniques to real-world data and contributing to informed decision-making processes.

R-Code :

```
##########NO. of Birth per month in new york city 1946-1959#

births <- scan("http://robjhyndman.com/tsdldata/data/nybirths.dat")

birthstimeseries <- ts(births, frequency=12, start=c(1946,1))

birthstimeseries

plot.ts(birthstimeseries)


###One method of establishing the underlying trend (smoothing out peaks and troughs)

in a set of data is using the moving averages technique##


# Calculate the 12-month moving average

moving_average <- filter(birthstimeseries, filter= c(1/24, 1/12, 1/8, 1/6, 1/4, 1/3, 1/4, 1/6, 1/8, 1/12, 1/24), sides=2)


# Handle NA values introduced by the filter

moving_average[is.na(moving_average)] <- NA


# Plot the original time series

plot.ts(birthstimeseries, main="Number of Births per Month in NYC (1946-1959) with 12-Month Moving Average", ylab="Number of Births", xlab="Year", col="blue")


# Add the moving average to the plot

lines(moving_average, col="red")
```

```
# Add a legend

legend("topright", legend=c("Original Data", "12-Month Moving Average"), col=c("blue", "red"), lty=1)


## Fit a linear model to the time series

time_index <- time(birthstimeseries)

linear_model <- lm(birthstimeseries ~ time_index)


# Get the fitted values (trend component)

trend <- fitted(linear_model)

# Detrend the time series by subtracting the trend component

detrended_series <- birthstimeseries - trend

# Plot the original time series

plot.ts(birthstimeseries, main="Original and Detrended Time Series", ylab="Number of Births", xlab="Year", col="blue")

lines(trend, col="red")


# Plot the detrended series

plot.ts(detrended_series, main="Detrended Time Series", ylab="Detrended Number of Births", xlab="Year", col="green")


# Add legends to both plots

legend("topright", legend=c("Original Data", "Trend"), col=c("blue", "red"), lty=1)

legend("topright", legend=c("Detrended Series"), col=c("green"), lty=1)

#Now remove seasonality
```

```
# Calculate the mean for each month across all years

monthly_means <- tapply(birthstimeseries, cycle(birthstimeseries), mean)

monthly_means

# Replicate the monthly means to match the length of the original time series

monthly_means_rep <- rep(monthly_means, length(birthstimeseries) / 12)

monthly_means_rep

# Subtract the monthly means from the original time series to get the deseasonalized series

deseasonalized_series <- birthstimeseries - monthly_means_rep

deseasonalized_series

# Plot the original time series

plot.ts(birthstimeseries, main="Original and Deseasonalized Time Series", ylab="Number of Births", xlab="Year", col="blue")

# Plot the deseasonalized series

lines(deseasonalized_series, col="red")

# Add a legend

legend("topright", legend=c("Original Data", "Deseasonalized Series"), col=c("blue", "red"), lty=1)

# Fit a linear model to the deseasonalized series to estimate the trend

time_index <- time(deseasonalized_series)

trend_model <- lm(deseasonalized_series ~ time_index)

# Get the fitted values (trend component)

trend_component <- fitted(trend_model)

# Subtract the trend component to get the residuals

residuals <- deseasonalized_series - trend_component

residuals
```

```
# Plot the trend component on the deseasonalized series plot

plot.ts(deseasonalized_series, main="Deseasonalized Series and Trend Component",
ylab="Deseasonalized Number of Births", xlab="Year", col="red")

lines(trend_component, col="blue")

legend("topright", legend=c("Deseasonalized Series", "Trend Component"), col=c("red", "blue"),
lty=1)

# Plot the residuals

plot.ts(residuals, main="Residuals (Deseasonalized Series - Trend)", ylab="Residuals",
xlab="Year", col="green")

# Analyze the residuals

acf(residuals, main="ACF of Residuals")

pacf(residuals, main="PACF of Residuals")

hist(residuals)

boxplot(residuals)

# Add a legend

legend("topright", legend=c("Residuals"), col=c("green"), lty=1)

#NOW APPLY LOG TRANSFORMATION METHOD

# Apply the log transformation to stabilize variance

log_birthstimeseries <- log(birthstimeseries)

log_birthstimeseries

# Calculate the mean for each month across all years

monthly_means_log <- tapply(log_birthstimeseries, cycle(log_birthstimeseries), mean)

# Replicate the monthly means to match the length of the original time series

monthly_means_rep_log <- rep(monthly_means_log, length(log_birthstimeseries) / 12)

monthly_means_rep_log
```

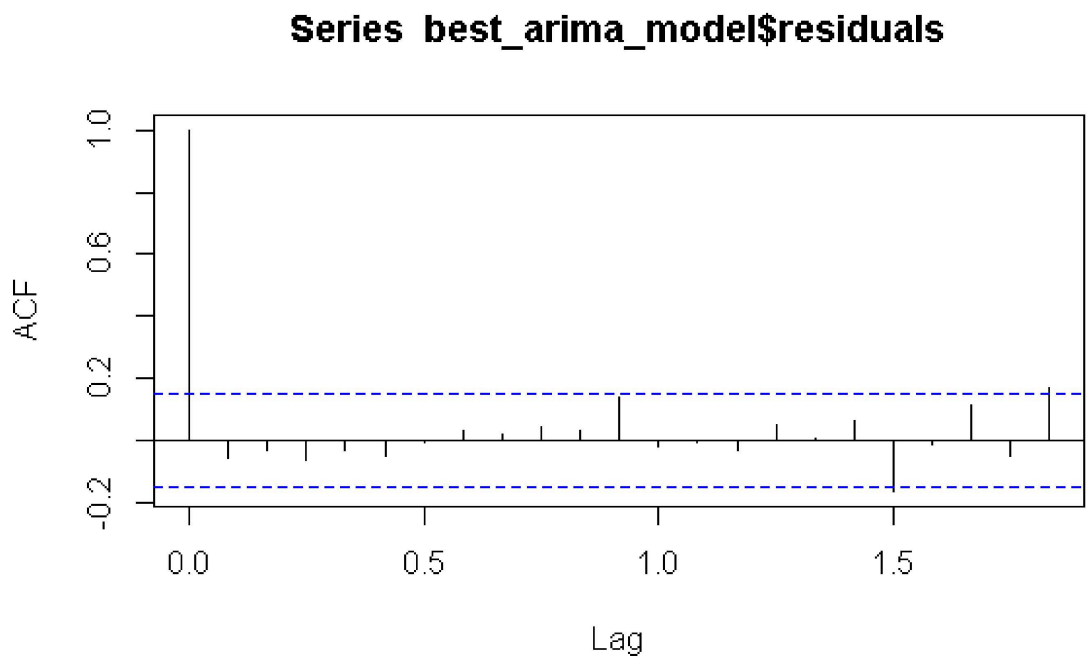
```
# Subtract the monthly means from the log-transformed time series to get the deseasonalized series

deseasonalized_log_series <- log_birthstimeseries - monthly_means_rep_log

deseasonalized_log_series


# Fit a linear model to the deseasonalized series to estimate the trend

time_index <- time(deseasonalized_log_series)

trend_model_log <- lm(deseasonalized_log_series ~ time_index)


# Get the fitted values (trend component)

trend_component_log <- fitted(trend_model_log)


# Subtract the trend component to get the residuals

residuals_log <- deseasonalized_log_series - trend_component_log

residuals_log


# Plot the original log-transformed time series

plot.ts(log_birthstimeseries, main="Original and Deseasonalized Log-transformed Time Series", ylab="Log(Number of Births)", xlab="Year", col="blue")


# Plot the deseasonalized log-transformed series

lines(deseasonalized_log_series, col="red")


# Add a legend
```

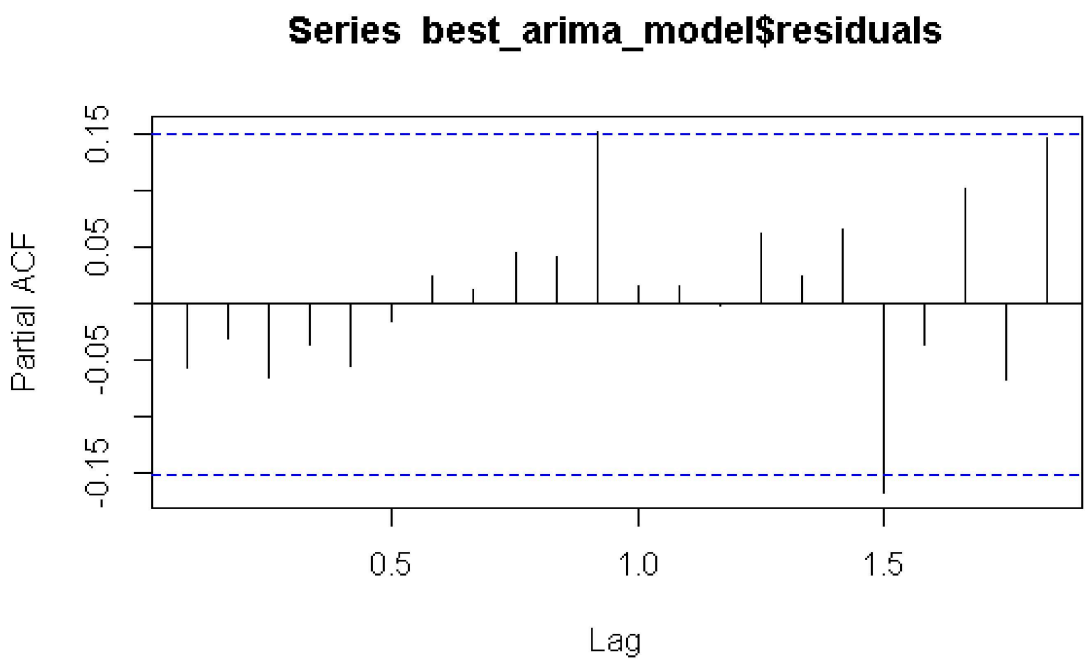
```
legend("topright", legend=c("Original Log-transformed Data", "Deseasonalized Series"), col=c("blue", "red"), lty=1)


# Plot the deseasonalized series with the trend component

plot.ts(deseasonalized_log_series, main="Deseasonalized Log-transformed Series and Trend Component", ylab="Deseasonalized Log(Number of Births)", xlab="Year", col="red")

lines(trend_component_log, col="blue")

legend("topright", legend=c("Deseasonalized Series", "Trend Component"), col=c("red", "blue"), lty=1)


# Plot the residuals

plot.ts(residuals_log, main="Residuals (Deseasonalized Log-transformed Series - Trend)", ylab="Residuals", xlab="Year", col="green")


# Analyze the residuals

acf(residuals_log, main="ACF of Residuals")

pacf(residuals_log, main="PACF of Residuals")

hist(residuals_log)

boxplot(residuals_log)

# Add a legend

legend("topright", legend=c("Residuals"), col=c("green"), lty=1)


#Load the necessary library

library(forecast)


# Automatically select the best ARIMA model
```

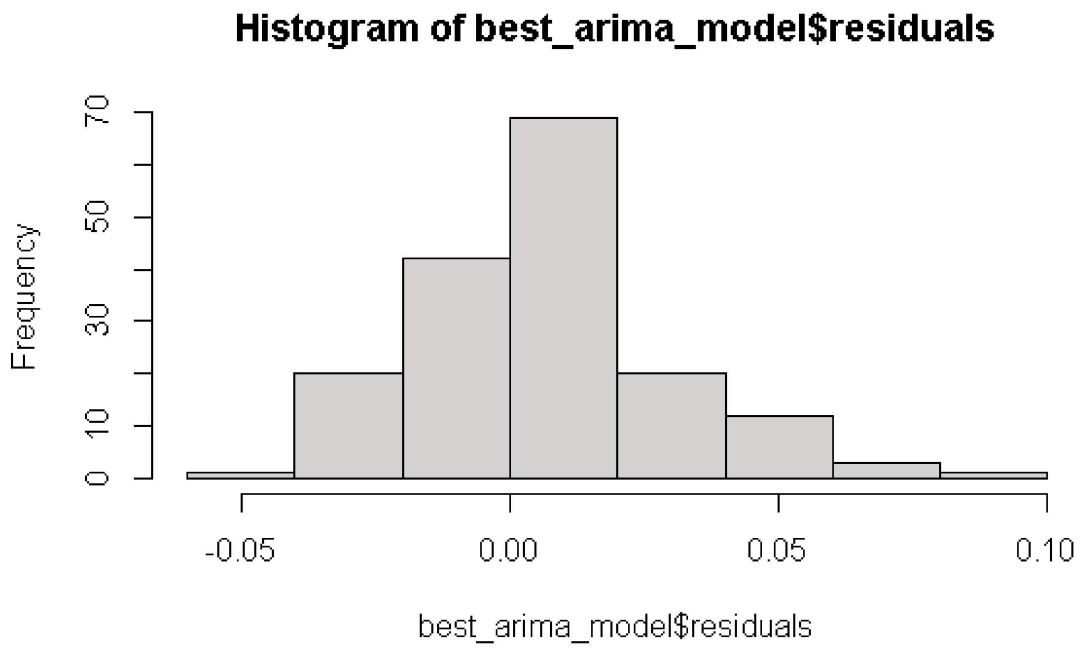
```
best_arima_model <- auto.arima(log_birthstimeseries)

# Display the summary of the best ARIMA model

summary(best_arima_model)

# Forecast using the best ARIMA model for the next 24 months

arima_forecast <- forecast(best_arima_model, h=24)

# Plot the original log-transformed series and the forecast

plot(arima_forecast, main="ARIMA Forecast with Log-transformed Data", ylab="Log(Number of Births)", xlab="Year")

# Add the original data to the plot

lines(log_birthstimeseries, col="blue")

# Add a legend

legend("topright", legend=c("Original Series", "Forecast"), col=c("blue", "red"), lty=1)

#Analyze

ts.plot(best_arima_model$residuals)

acf(best_arima_model$residuals)

pacf(best_arima_model$residuals)

pacf(best_arima_model$residuals,40)

acf(best_arima_model$residuals,40)
```

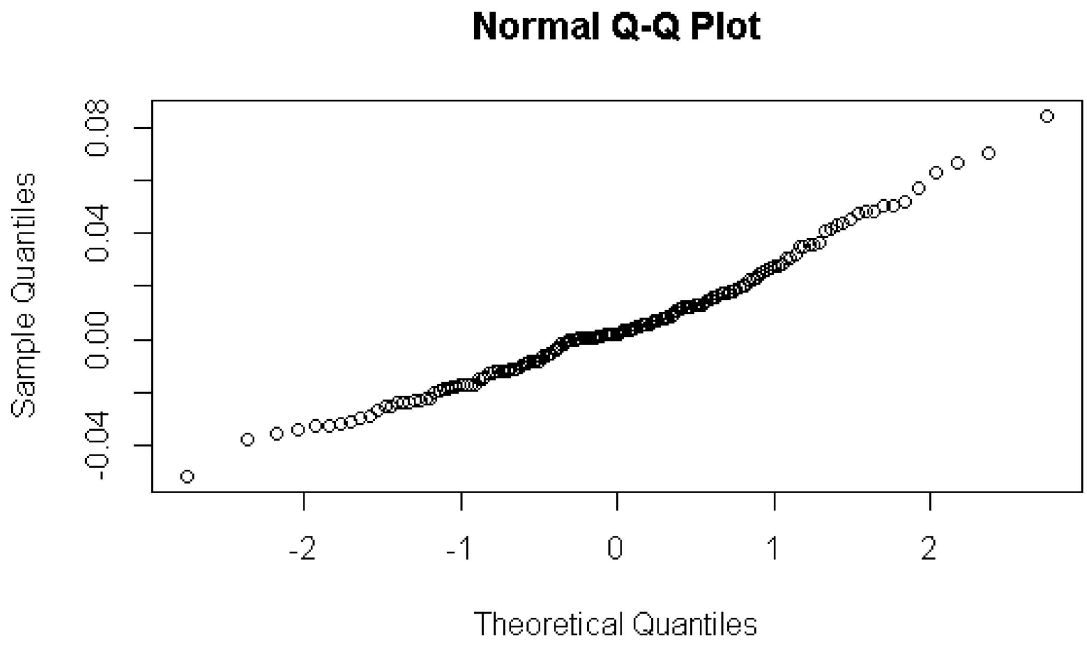
```
hist(best_arima_model$residuals)

qqnorm(best_arima_model$residuals)
```

Output:




Series best_arima_model$residuals
Partial ACF
0.15
0.05
-0.05
-0.15
0.5
1.0
1.5
Lag


Histogram of best_arima_model$residuals
Frequency
70
50
30
10
0
-0.05
0.00
0.05
0.10
best_arima_model$residuals



Responsibilities and Tasks

During my internship, I was assigned a range of responsibilities that provided me with hands-on experience in time series analysis and data analytics. My main responsibilities included:

Data Collection and Preparation:

- Acquiring the dataset of monthly birth counts in New York City from January 1946 to December 1959.
- Cleaning and preprocessing the data to ensure accuracy and consistency for analysis.

Exploratory Data Analysis (EDA):

- Visualizing the time series data to understand its structure and identify initial patterns.
- Conducting summary statistics to get an overview of the data distribution and characteristics.

Time Series Decomposition:

- Decomposing the time series into its components: trend, seasonality, and irregular components.
- Applying moving averages to smooth the data and highlight long-term trends.

Trend and Seasonality Analysis:

- Fitting linear models to the time series to estimate and remove trends.
- Calculating and removing seasonal effects to deseasonalize the series.

Transformation and Detrending:

- Applying log transformations to stabilize variance.
- Detrending the time series by subtracting the estimated trend component.

Model Building and Forecasting:

- Using ARIMA models to capture the dynamics of the time series and make future forecasts.
- Evaluating different ARIMA model configurations and selecting the best fit using auto.arima.

Residual Analysis:

- Analyzing the residuals of the fitted models to ensure they resemble white noise.
- Using ACF and PACF plots, histograms, and Q-Q plots to assess the residuals.

Reporting and Documentation:

- Documenting the entire analysis process, methodologies, and findings.
- Preparing visualizations and summary reports to communicate insights effectively.

By carrying out these responsibilities and tasks, I was able to gain a thorough understanding of time series analysis, improve my data analysis skills, and contribute valuable insights to the this problem.

Skill Developed

During the course of my internship, I developed a diverse set of skills that enhanced my proficiency in time series analysis, data manipulation, and statistical modeling. These skills are broadly categorized as follows:

Technical Skills

Time Series Analysis:

- **Decomposition**: Gained expertise in breaking down time series data into trend, seasonal, and irregular components using both visual and statistical methods.
- **Trend and Seasonality Detection**: Improved ability to identify and analyze trends and seasonal patterns within time series data.
- **Moving Averages**: Learned to apply moving averages to smooth data and highlight underlying trends.

Statistical Modeling:

- **Linear Regression**: Enhanced skills in fitting linear models to time series data to estimate and remove trend components.
- **ARIMA Modeling**: Gained hands-on experience in building and validating ARIMA models for forecasting future values of time series data.
- **Residual Analysis**: Developed the ability to analyze and interpret model residuals to ensure model adequacy and performance.

Data Transformation:

- **Log Transformation**: Acquired knowledge of applying log transformations to stabilize the variance in time series data.
- **Deseasonalization**: Learned to remove seasonal effects from time series data to obtain a deseasonalized series.

Forecasting:

- **Model Selection**: Improved skills in selecting the best-fit model for forecasting using automated model selection tools like auto.arima.
- **Forecast Interpretation**: Enhanced ability to interpret and visualize forecast results, including confidence intervals and prediction intervals.

Data Visualization:

- **Plotting Techniques**: Developed proficiency in creating various plots (time series plots, ACF/PACF plots, histograms, Q-Q plots) to visualize data and model results.
- **Graphical Presentation**: Improved skills in presenting data and findings through clear and informative visualizations.

Analytical Skills

Exploratory Data Analysis (EDA):

- Gained expertise in conducting EDA to understand data distributions, identify patterns, and detect anomalies.
- Enhanced ability to summarize data using descriptive statistics and visual tools.

Problem-Solving:

- Improved problem-solving skills by tackling challenges in time series analysis, such as dealing with missing values, outliers, and complex seasonal patterns.
- Developed critical thinking abilities to choose appropriate methods and techniques for different stages of the analysis.

Programming Skills

R Programming:

- Strengthened proficiency in using R for data analysis, time series manipulation, and statistical modeling.
- Learned to use various R packages (e.g., forecast, tseries) for implementing advanced analytical techniques.

Code Documentation:

- Enhanced ability to document code effectively, ensuring reproducibility and clarity for future reference.

Conclusion

The internship experience provided me with a comprehensive understanding of time series analysis and its practical applications in real-world scenarios. Throughout the problem, I had the opportunity to engage with various stages of data analysis, from data collection and preprocessing to advanced statistical modeling and forecasting.

Overall, this internship has been a valuable learning experience, providing me with practical skills and knowledge that are crucial for a career in data analysis and statistical modeling. This problem not only reinforced my theoretical understanding of time series analysis but also allowed me to apply these concepts to real-world data, resulting in meaningful insights and forecasts.

I am grateful for the guidance and support provided by my supervisor. This internship has equipped me with the confidence and expertise to tackle future challenges in the field of time series analysis, and I look forward to applying these skills in my professional journey.