

Time Series Analysis

northeastern university

aly6015

Joseph Manseau

by

abhijeet sudhakar, dhanashree chavan, okan yazmacilar, preeti bhole

date

9th May, 2020

Table of Contents

[Introduction 3](#_Toc40029606)

[Part A - Exploratory Data Analysis 3](#_Toc40029607)

[Part B – Time Series Decomposition 7](#_Toc40029608)

[Part C – Check for Autocorrelation 8](#_Toc40029609)

[Part D – Fit a Model 10](#_Toc40029610)

[Part E – Determine Forecasts 12](#_Toc40029611)

[Appendix A – References 16](#_Toc40029612)

# Introduction

The objective of this analysis and modelling is to review time series theory with R packages. We will be following a modelling procedure on the **nottem** (A time series object containing average air temperatures at Nottingham Castle in degrees Fahrenheit for 20 years) dataset as follows:

1. Perform exploratory data analysis
2. Decomposition of data
3. Check for Autocorrelation
4. Fit a model
5. Determine forecasts

# Part A - Exploratory Data Analysis

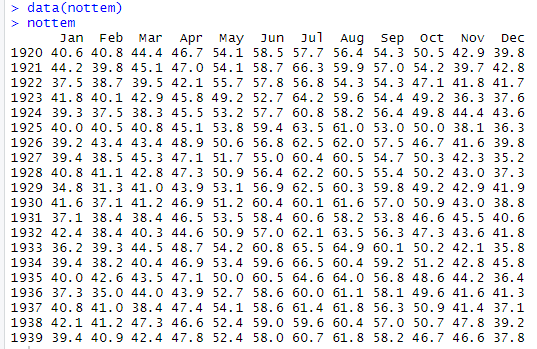
Install the required packages



Load the packages



Load the “nottem” data into R



The nottem dataset in R provides average air temperatures at Nottingham Castle in degrees Fahrenheit from 1920 to 1939. Let’s check the class of this dataset:



Since this dataset is already a time series, no conversion is required.

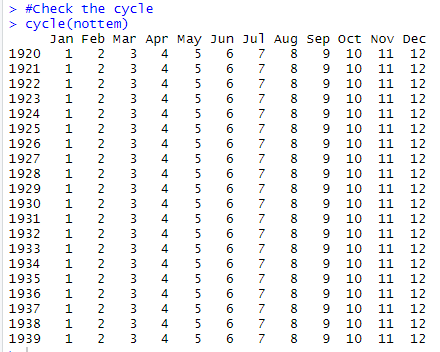
Now, we will review the summary statistics of this dataset and plot some graphs in R.



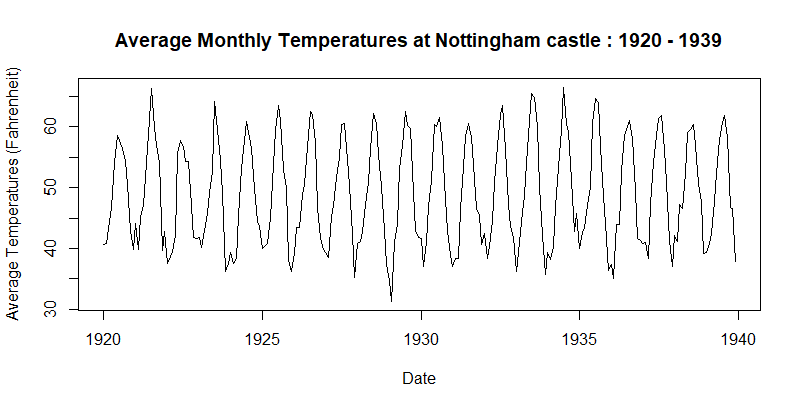
We can see that it is a monthly data.



There are no missing values. Thus, no special handling is required to remove them.

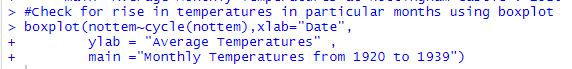


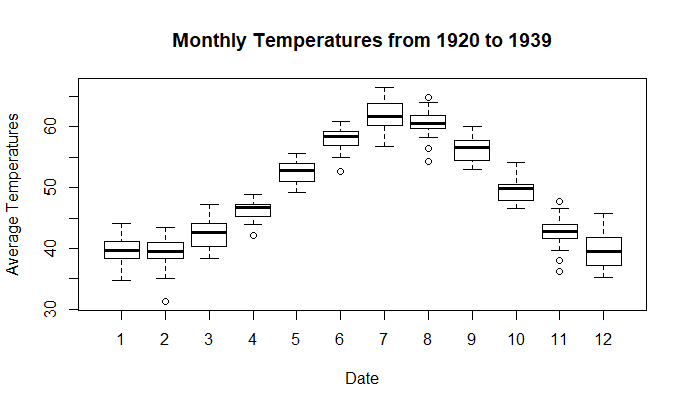




We can see from this time series that there is a seasonal variation in average temperatures on a monthly basis. There seems to be a peak in temperatures in the middle of the year and then drops towards the end of the year. This time series could be described as an additive model as the seasonal fluctuations are nearly constant, there is no particular trend as such since it is almost the same for 20 years, and the random variations also seem to be nearly constant in size over the period.

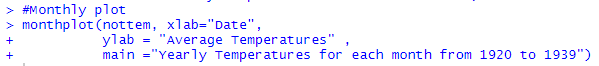
Let’s check if there are rise in temperatures in particular months over these years using boxplot:

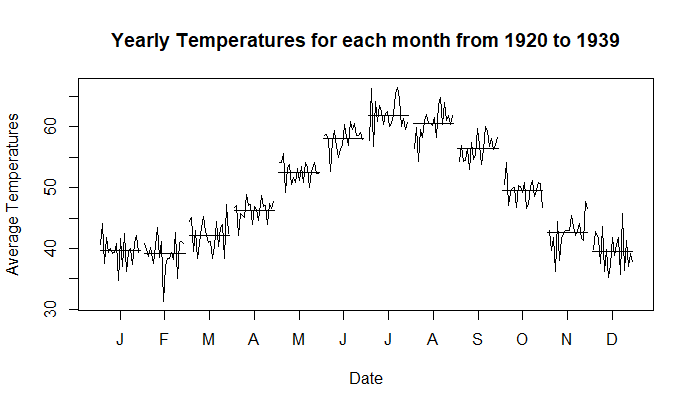




We can make some inferences using above boxplot:

1. There is an increase in average temperatures starting from January till July followed by subsequent decrease till December.
2. Average temperatures in December, January and February remain mostly constant to around 40 degree fahrenheit.
3. There seems to be some extreme temperature variations in the months of February, April, June, August and November for a few of the years. We can confirm this by plotting the dataset using the monthplot() function of R.



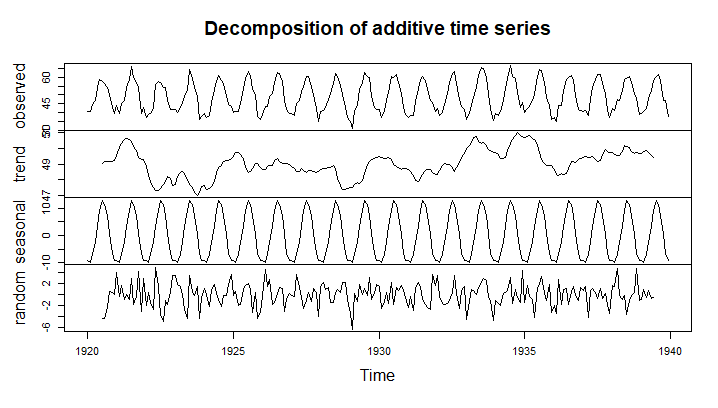


As confirmed, we can clearly visualize the peaks in temperatures in a few months for some years. Now, we will decompose the data into its components in the next section.

# Part B – Time Series Decomposition

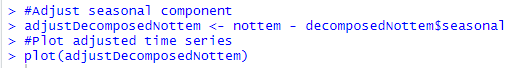
We would decompose the “nottem” dataset into trend, seasonal and random components using the decompose() function in R.

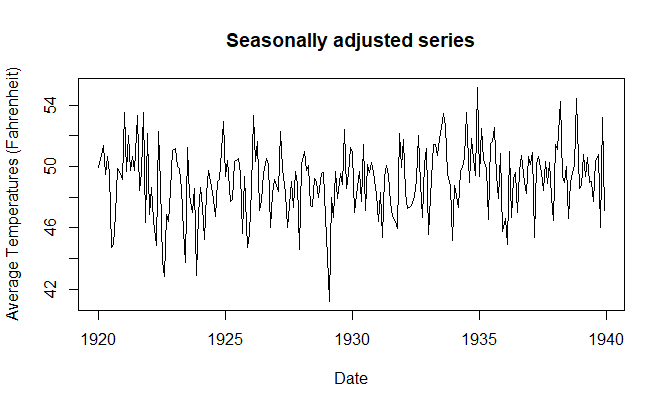




We can see that the time series is decomposed into three components: trend, seasonal and random. The trend component seems to remain almost the same with some peaks and drops in a few years. The seasonal component remains the same across each year and depicts rise in temperatures in the middle of the year.

Let’s seasonally adjust the time series by estimating the seasonal component and then subtracting the seasonal component from the original time series. Then, we will plot this seasonally adjusted series.



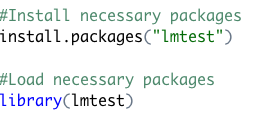


We can see that after removing seasonal components, random and trend components remain. The irregularities are fairly the same in size over the period.

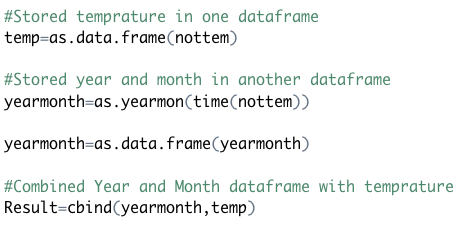
# Part C – Check for Autocorrelation

Before creating the model for time series analysis, it is crucial to check the auto-corelation in the model. Hence we performed the Durbin-Watson hypothesis test to check auto-corelation.

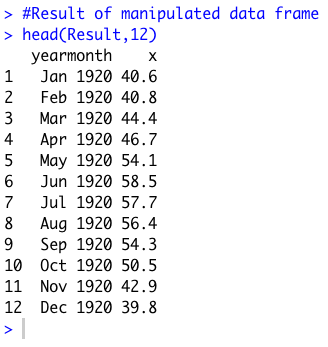
First loaded the required package for the test as follows:



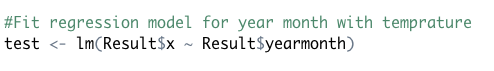
To perform the Durbin-Watson hypothesis test, first we need to perform linear regression of the time with the temperature. Hence we first manipulated the data frames to be suitable for linear regression as follows:



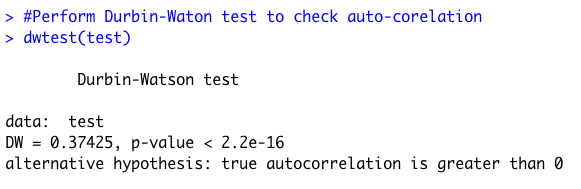
Below is the screenshot of the manipulated data frame to be used for regression.



Once the data frame was manipulated to fit the regression model as performed linear regression as follows:

****

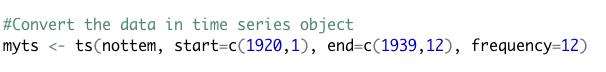
Next we performed the Durbin-Watson hypothesis test to check auto-corelation in the residuals of our regression analysis.



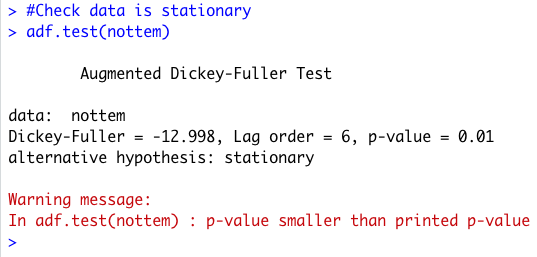
As p-value is less than the significance level 0.05, hence our model has auto-correlation and is appropriate to perform further analysis.

# Part D – Fit a Model

In order to perform time series analysis, we first need to convert our data to time series object which we did as follows:



Moreover, before building the model, we need to check whether our data is stationary hence we used Augmented Dickey-Fuller Test to verify this as follows:



As the p-value is less than the significance level 0f 0.05, hence we can conclude that our data is stationary and appropriate to perform the time series forecasting.

We performed the time series analysis using the ARIMA (Autoregressive Integrated Moving Average) model. This model is used to perform univariate time series analysis.

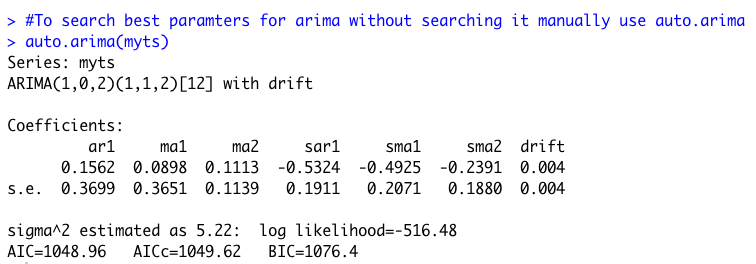
But before fitting the model we need to find the values of p, d and q which stands for:

p = number of auto-regressive terms.

d= number of nonseasonal differences needed for stationarity.

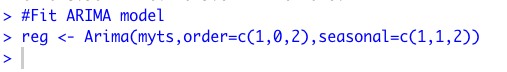
q= number of lagged forecast errors in the prediction equation.

To find the best parameter values for the ARIMA model we used the auto.arima function instead of calculating it manually. The auto.arima function returns the best ARIMA model according to either AIC, AICc or BIC value.

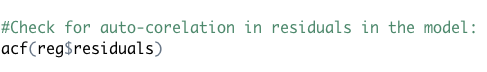


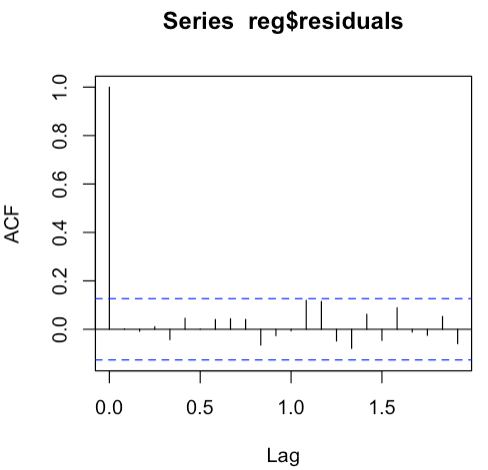
From the above function we got our values of p, d, q as (1,0,2). Also the seasonal model has an autoregressive term of first lag at model period 12 units, in our case months.

Now we will fit our model.



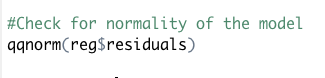
After fitting the model we checked whether there is auto-correlation in the residuals in the model through the function ACF as follows:

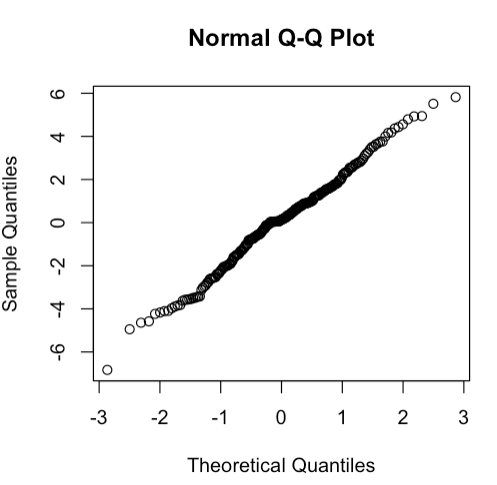




From the above acf plot we can see that all values are lying under the blue-line hence there is no auto-corelation in the residuals of the model.

We also checked the normality of the model by plotting the qqplot.

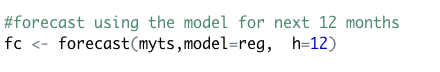




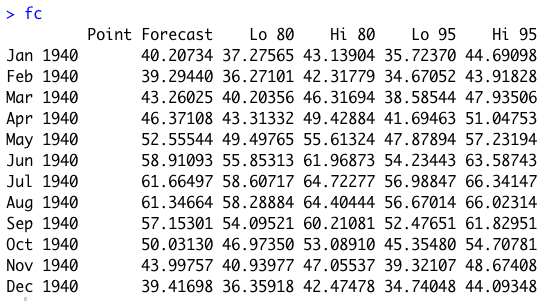
From the qqplot, we can see that our model has normality as well.

# Part E – Determine Forecasts

Finally we can plot a forecast of the time series using the forecast function, again from the forecast R package, with a 80% and 95% confidence interval where h is the forecast horizon period in months. We predicted temperature for next 12 months after 1939, we also passed our time series object and our model ‘reg’ as follows:

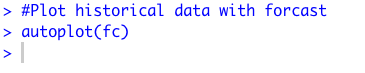


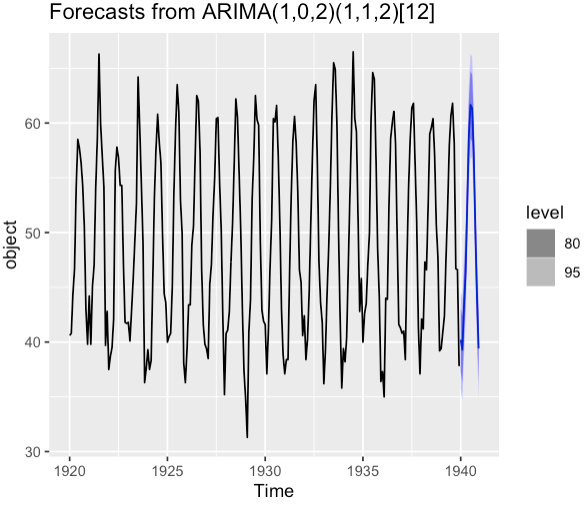
Below is the forecast of temperature for next 12 months.



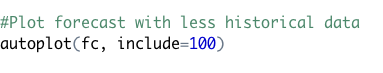
Point forecasts are the best way to show the predicted values. But sometimes it is difficult to rely on them, hence we can show how far off these point forecasts will be by the 80% and 95% confidence interval.

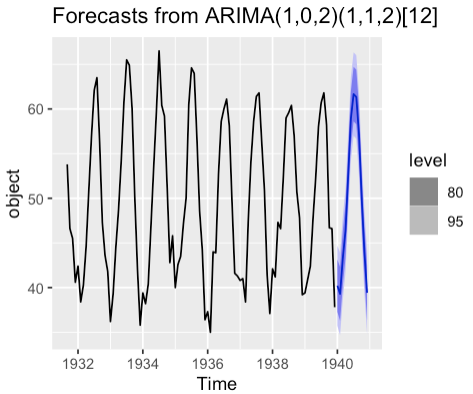
Next we plotted the historical data with the forecasted values as follows:





Sometimes it is difficult to view the forecast due to earlier data, hence we excluded some of the earlier data and plotted the graph once again as follows:





Hence we can see that the data seems to be stationary in the future as well and its capturing the seasonal patterns also.

**Conclusion**

We opened this paper noting the importance of time series analysis. We did necessary installations and loads. In our data there wasn't any particular trend. And when we check the data monthly, we see that there are some extreme temperature variations in the months of February, April, June, August and November for a few of the years. After we decompose the data, the trend and seasonal component remain almost the same. In the autocorrelation part first we perform linear regression and Durbin-Watson hypothesis test. Our P-value was less than 0.05 and it was okay to continue. We performed the ARIMA model using the auto.arima function and then we fit the model. We predicted the temperature for the next 12 months after 1939.

# Appendix A – References

● Kimnewzealand (Sept, 2017), *An Introduction To Grid Search*. Retrieved from http://rstudio-pubs- static.s3.amazonaws.com/311446\_08b00d63cc794e158b1f4763eb70d43a.html

● *Forecast*. Retrieved from <https://www.rdocumentation.org/packages/forecast/versions/8.12/topics/auto.arima>