Time Series Analysis of Monthly Milk Production

PROJECT REPORT

Table of Content

Introduction

What is time series?

Series of observations recorded sequentially over a period of time (i.e. A collection of observations recorded along with the time)

We have taken our data of monthly milk production per pound. We are going to apply all the basic methods for checking different aspects of Time Series data like covariance stationarity, model identification, parameter estimation etc.

Data Set

The data contains data of monthly Milk production in pound per cow for 14 years (from January1962 to December -1975).

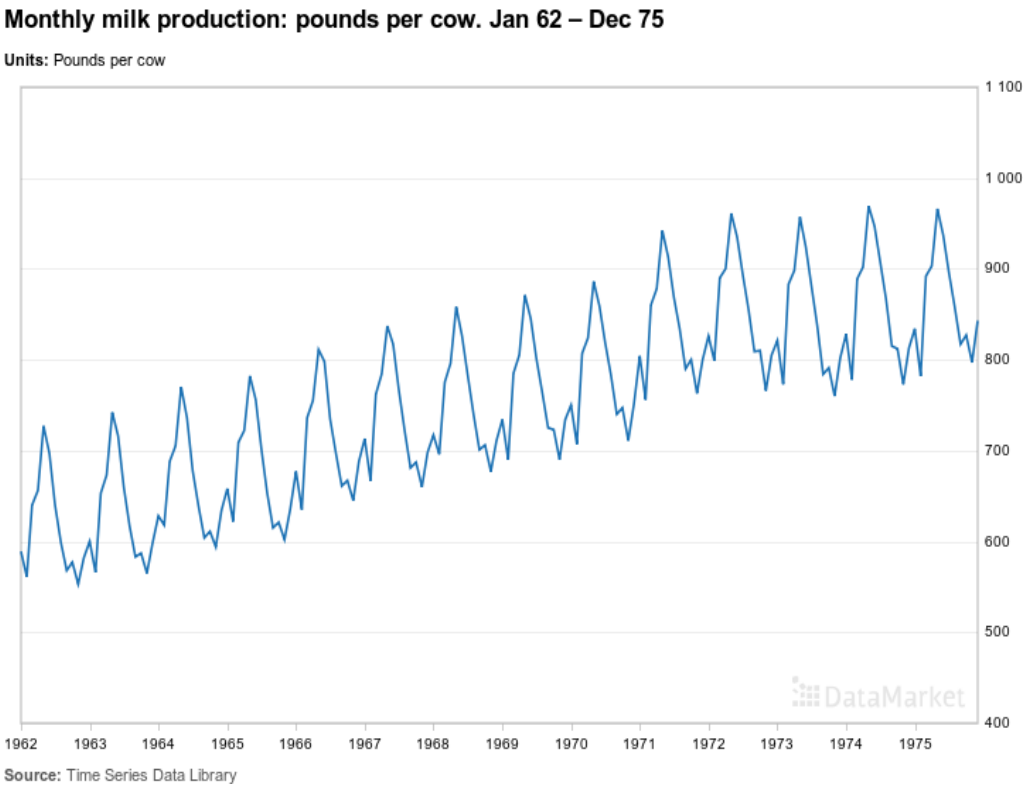
Head and Tail of Data

Date	Quantity	Date	Quantity
1962-01	589	.	.
1962-02	561	.	.
1962-03	640	.	.
1962-04	656		
1962-05	727	1974-08	867
1962-06	697	1974-09	815
1962-07	640	1974-10	812
1962-08	599	1974-11	773
1962-09	568	1974-12	813
1962-10	577	1975-01	834
1962-11	553	1975-02	782
1962-12	582	1975-03	892
1963-01	600	1975-04	903
1963-02	566	1975-05	966
1963-03	653	1975-06	937
1963-04	673	1975-07	896
.	.	1975-08	858
.	.	1975-09	817
.	.	1975-10	827
.	.	1975-11	797
.	.	1975-12	843

Detailed information

Dataset title	Monthly milk production: pounds per cow. Jan 62 – Dec 75
Provider	Time Series Data Library
Source URL	http://datamarket.com/data/list/?q=provider:tsdl
Units	Pounds per cow
Dataset metrics	168 fact values in 1 time series.
Time granularity	Month
Time range	Jan 1962 – Dec 1975

Data Plot



Summary

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
553.0	677.8	761.0	754.7	824.5	969.0

Our data is a discrete-point data, taken at successive integer points in time.

Our objective would be:

1. Is the data purely random or there is some deterministic component present in it (like trend and seasonal components)?
2. To choose the proper model.
3. To forecast future values based on the given time series data and to check how close it is to the real data.

Analysis

1. Randomness Test:

Firstly, we have checked whether our data is purely random or not. And for that we have performed the "turning point" non-parametric test (a statistical test of the independence of a series of random variables).
We construct our hypothesis-

$$H_0: The\ data\ is\ purely\ random$$
$$against$$
$$H_1: The\ data\ is\ not\ purely\ random$$

Y_i is a turning point if $Y_i > Y_{i-1}$ and $Y_i > Y_{i+1}$ or $Y_i < Y_{i-1}$ and $Y_i < Y_{i+1}$.

Yi=1 if it is a turning point else 0
D=$\sum_{i=2}^{n-1} Yi$

Where D is the total number of turning points.

Under H_0 –

$$E(D) = \frac{2}{3}(n-2)$$
$$Var(D) = \frac{(16n-29)}{90}$$

$$z = \frac{D-E(D)}{\sqrt{var(D)}} \sim N(0,1) \text{ under} H_0$$

In our case, n= 168, D= 79

Calculated |z|= 5.83, tabulated z(5% level of significance)= **1.96**

i.e. **cal(|z|) ≥ tab(z**), so we ***reject*** our null hypothesis at 5 % level of significance. Thus, our data is not random and there is some trend in the model. Now, we will proceed towards checking trend in the model.

Testing for the presence of trend

Relative ordering Test

Our hypothesis of interest is:

$$H_0: there\ is\ no\ trend\ in\ the\ data$$
$$against$$
$$H_1: trend\ is\ present\ in\ the\ data.$$

We define discordances as following:

$$q_{ij} = \begin{cases} 1\ , & y_i > y_j, where\ i < j \\ 0, & otherwise \end{cases}$$

$$Q = \sum_{i<j} \sum q_{ij} : \quad number\ of\ \ discordant\ \ points\ in\ time\ series$$

$$E(Q) = \frac{n(n-1)}{4}$$

Kendall's rank correlation coefficient is defined as-

$$\tau = 1 - \frac{4Q}{n(n-1)} = 0.60878$$

$$Under\ H_0, \qquad E(\tau) = 0, v(\tau) = \frac{2(2n+5)}{9n(n-1)}$$

$$z = \frac{\tau - E(\tau)}{\sqrt{v(\tau)}} \sim N(0,1)$$

Calculated **|z|= 11.71392**, tab **|Z|= 1.96** (at 5 % level of significance)

i.e. **cal |z|>tab|z|** , hence we ***reject*** the null hypothesis at 5% level of significance .

Therefore, there is trend in data.

Test of presence of Seasonality

Friedman's Test

If d is the period of seasonality

$$S_{t-d} = S_t = S_{t+d}$$

Our hypothesis of interest –

$$H_0: There\ is\ no\ seasonality$$
$$against$$
$$H_1: H_0\ is\ not\ true.$$

Give the rank to all for a particular month and for a particular year.
Our data is 14 year data. First we find rank corresponding to each month and years.

$$M_{ij} : \underset{c}{rank}\ corresponding\ to\ i^{th}\ month\ and\ j^{th}\ year.$$

$$M_i = \sum_{j=1} M_{ij}\ , i = 1, \dots, 12\ , \quad E(M_i) = c\frac{(r+1)}{2}\ ,$$

$$Test\ statistics = \frac{\sum_{i=1}^{12(r)}(M_i - E(M_i)}{c(r+1)} \sim \chi^2_{r-1}$$

In our data –

$r = 12,\ c(r+1) = 182\quad \chi^2_{obs=S=149.9698}$

Tabulated value at degree of freedom (r-1): $\chi^2_{tab=19.67513757}$

since $^2_{obs} > {}^2_{tab}$, we **reject** our null hypothesis i.e

Seasonality is present in our data.

Determination of period of seasonality

By Differencing

We apply method of differencing to obtain the period of seasonality. By applying appropriate order of differencing, we can de-seasonalize the data.
Let's consider the model-

$$y_t = m_t + s_t + \in_t$$

If d is the period of seasonality, then-

$$S_{t-d} = S_t = S_{t+d}, \qquad \sum_k S_k = 0$$

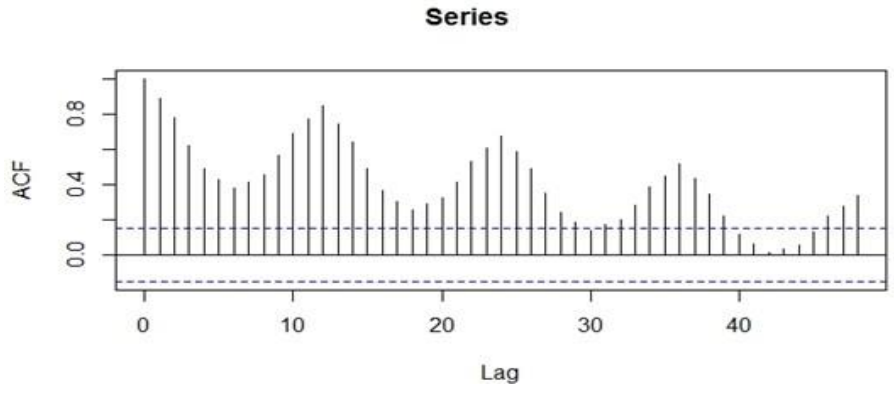
$$\mathrm{A}_d y_t = m_t - m_{t-d} + \in_t - \in_{t-d}$$

and the new time series will be -

$$\mathrm{A}_d y_t$$

If Friedman's non-parametric test suggests absence of seasonality in differenced data series then, d is the period of seasonality. Therefore, the new time series will be $\nabla_d Y_t$. Hence, our next task is to test for the seasonality in the $\nabla_d Y_t$ and if the Friedman's nonparametric test for seasonality suggests the absence of seasonality then our conclusion will be that the chosen period of seasonality d is perfect.

From the graph, we observe a high positive correlation at lags of 12, 24, 36 and hence, assume that d=12.



Applying Friedman's test after differencing, we get-

$$\chi^2_{obs,11} = 2.71735, \quad and \ \chi^2_{tab,11} = 19.675$$

$\chi^2_{obs} > \chi^2_{tab}$, we accept our null hypothesis i.e.the seasonality is absent

Estimation and Elimination of the trend and seasonal components of the Time Series

Model: $Y_t = m_t + S_t + e_t$

where

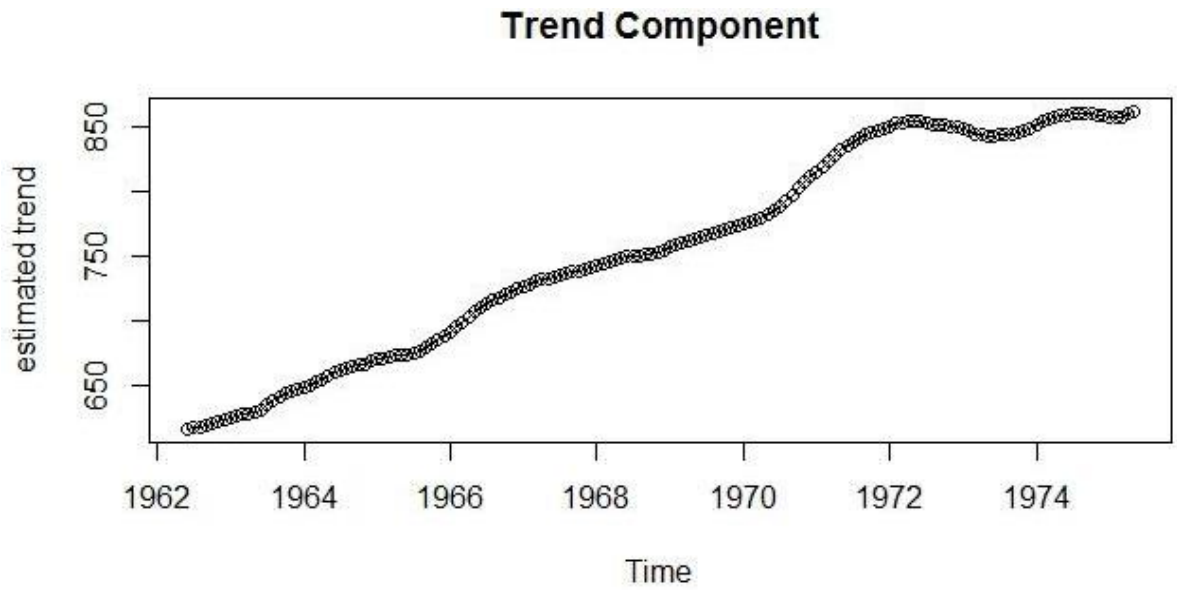
m_t : Trend component

S_t: Seasonal component

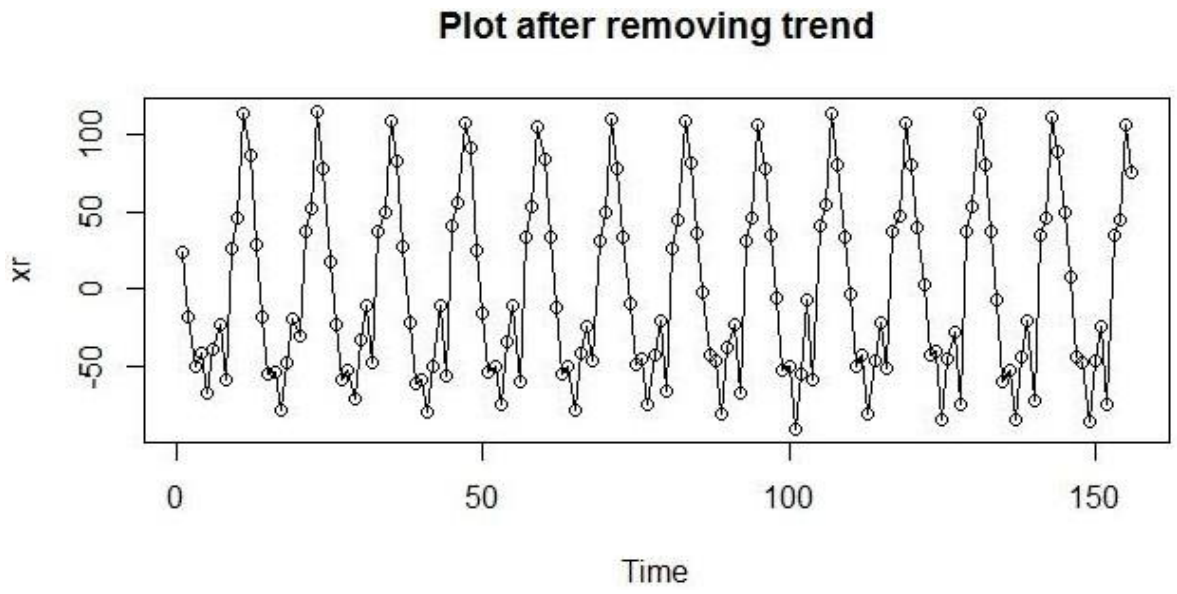
The model assumptions are $E(e_t) = 0$, $Var(e_t) = \sigma^2$, which is finite.

Based on this model we need to estimate the trend component and seasonal component and eventually eliminate them from the underlying data. We have used the Fast Changing Trend Method to do this. This method suggests that first we need to calculate the rough estimate of trend and using that estimates we need to de-trend the data and then we also have to de-seasonalize it.

Step 1: As we found out the period of seasonality is 12 then the corresponding rough estimate of trend m_t ,by the moving average filter is

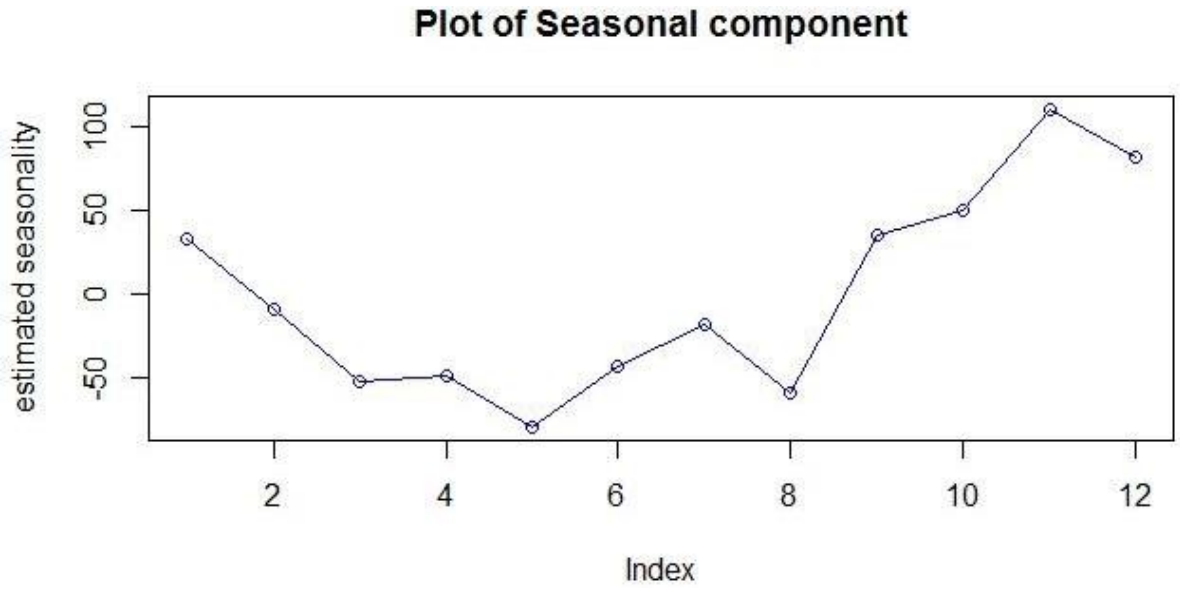


Step 2: De-trend the data as Yt- $\hat{mt}$



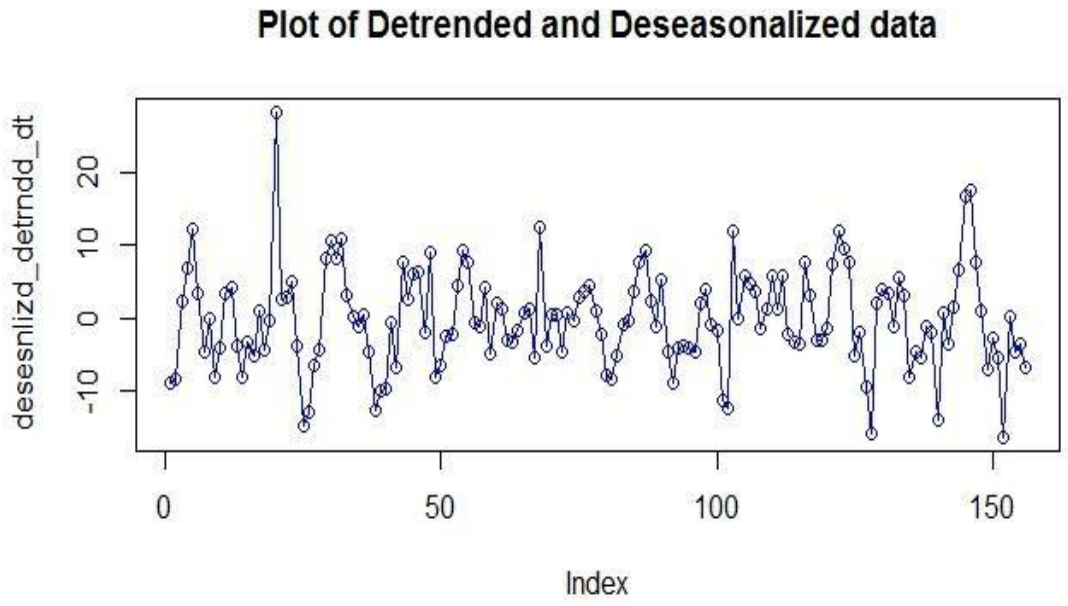
Step 3: Compute the average, say w_k of deviations $[y_{12(j-1)+k} - m_{12(j-1)+k}$: 7 ≤ 12(j-1)+k ≤ c-6 over the 14 years.

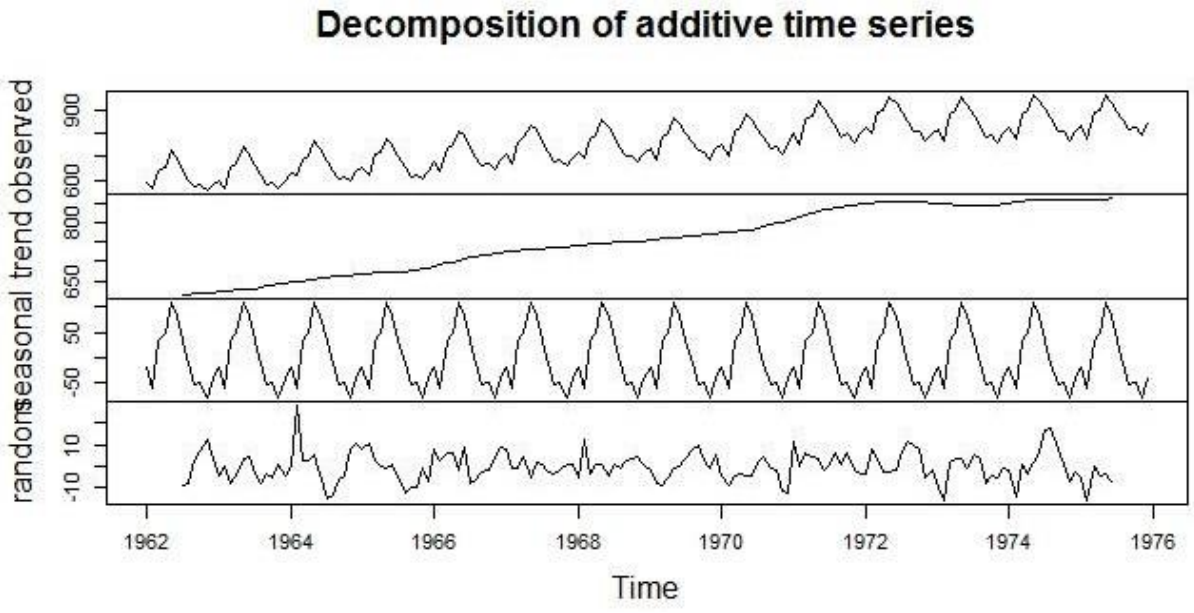
$S_k = w_k - (1/d) * \Sigma w_k$, k=1(1)12



Step 4: De-seasonalise the data and repeat the procedures until the trend as well as the seasonal component have been completely eliminated from the data.

Therefore, according to the test procedure we first de-trend and de-seasonalise our data and then recheck if the new time series data contains trend of seasonality. After performing the testing we find that the deterministic components are completely eliminated and we are left with the random error component, Z_t=Yt–$\hat{m_t}$-$s\hat{}_t$





Model Identification

Now to further use the data for forecasting we have to first model the data properly. Now, to check in which of the category does the data belong, we proceed in the following manner:

White Noise Checking:

Now, as we know that in case of a white noise data set ,the correlation function at lag h will be 0 if h≠0. So, we shall have to test the same for our dataset. Our hypothesis becomes:

$$H_0: \rho(h) = 0 \quad , \qquad against \quad H_1: \rho(h) \neq 0.$$

We know that -

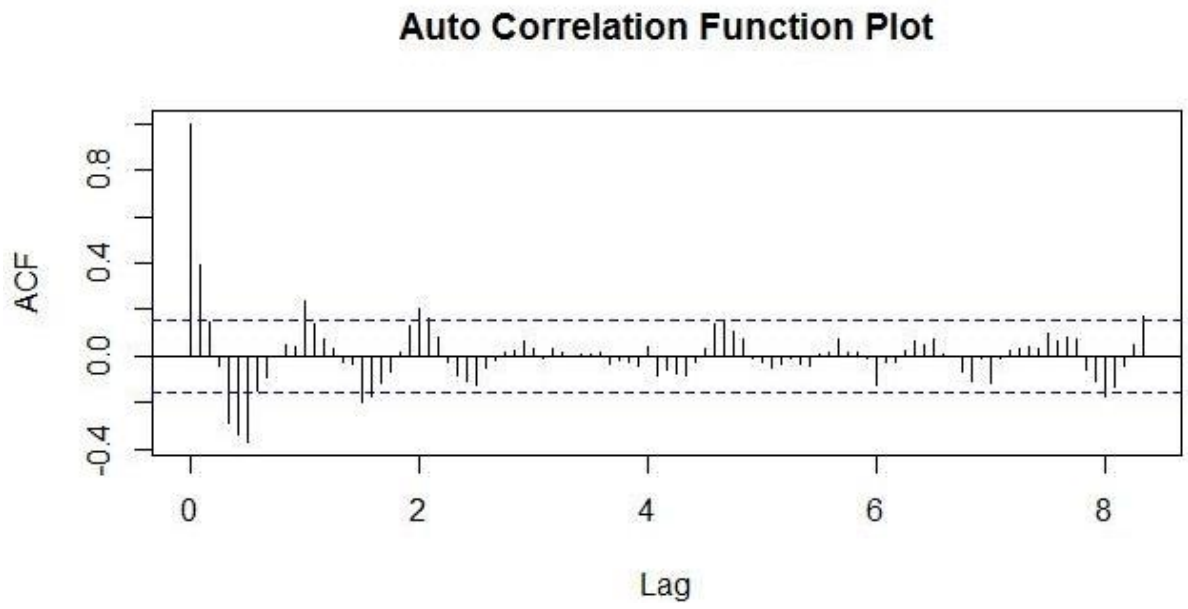
$z = \frac{\sqrt{n}(\hat{\rho}(h) - \rho(h))}{}$,

Follows N(0,1) asymptotically under the null hypothesis.

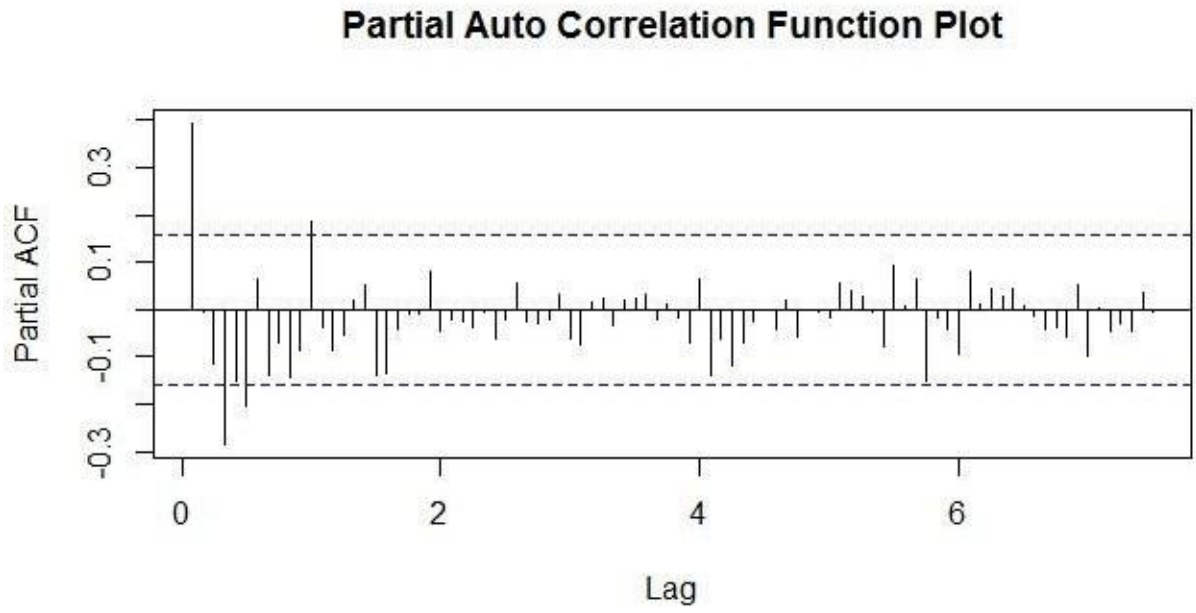
We calculated z = 4.916929, and z(tab)= 1.96, which is less than calculated z.
Hence, the data is not from a White Noise process. i.e. there is a significant correlation at lag 1.

Checking for Moving Average/Auto-Regressive/Auto-regressive moving average:

Furthermore, we can also check by plotting Auto Correlation Function (ACF) and observe that whether it tails off to 0 or not.
In order to find whether it is AR/MA/ARMA model, we will look at Auto-Correlation plot.
From the graph of ACF we find there is tailing off of spikes beyond the lag 8 and we assume it to be AR(p) with rough idea of having p around 8.



From The PACF plot, we observe that all the spikes lie well inside the acceptance band and tails off to 0 after 6 .Hence we identify our model to be an MA (q) Model with q lying around 6.



By both plot we find that it is ARMA (p, q) model .Assuming our model to be a stationary ARMA model, we proceed to estimate the order and coefficient of the

model and later incorporating the values of the coefficient in the model we cross-check our assumption of stationarity.

Estimation of model parameters and noise parameters

Now, to use the ARMA model we have to find the proper order of the model, in the sense that we will have to calculate the values of p and q.
From, the plot of ACF and PACF we have got a clue that the values of p and q will lie around 8 for p and around 6 for q. Now, fitting the ARMA model of order (p, q) we have calculated the value of the model Akaike Information Criteria (AIC). Now, the model for which the value of AIC is minimum considered as our final ARMA model.

	[, 1]	[, 2]	[, 3]	[, 4]	[, 5]	[, 6]	[, 7]	[, 8]	[, 9]
[1,]	1040.646	1019.682	1018.43	1014.551	1016.293	1008.863	981.8846	983.7921	985.048
[2,]	1016.038	1018.037	1019.235	1016.195	996.9002	994.5493	983.8104	985.503	986.34
[3,]	1018.036	1020.023	993.109	1003.478	988.747	989.9568	983.4321	985.4073	983.305
[4,]	1017.895	993.1132	994.6545	1004.967	987.1751	992.2961	984.9954	987.3395	984.675
[5,]	1006.032	993.937	991.9828	993.7346	992.8249	990.1397	988.2256	988.5518	990.779
[6,]	1004.05	993.1902	993.7215	993.001	984.5738	986.2331	985.8155	986.1012	986.165
[7,]	998.8695	998.6731	1000.671	994.2141	985.5086	981.0993	983.8539	985.7247	987.805
[8,]	999.9767	1000.673	993.0862	991.9938	993.7899	983.9238	985.7794	988.0408	989.857
[9,]	998.6118	992.6298	993.9975	993.7729	995.774	985.6872	987.6816	989.5835	990.697

We observe that, AIC is the least for ARMA(7,6) model (model AIC = 981.0993).
So, **estimated value of p = 7** and **estimated value of q= 6.**

ML Estimation of Parameters

We shall consider the Maximum Likelihood Estimation technique to calculate the value of the 15 (7 AR parameters+6 MA parameters+1 intercept+σ^2) parameters from the random data. The estimates of the parameters obtained as:

Coefficients

$\boldsymbol{\phi}_1$	$\boldsymbol{\phi}_2$	$\boldsymbol{\phi}_3$	$\boldsymbol{\phi}_4$	$\boldsymbol{\phi}_5$	$\boldsymbol{\phi}_6$	$\boldsymbol{\phi}_7$
-0.1543	0.256904	-0.2199	-0.23248	0.411716	-0.24344	-0.10272
θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	$\boldsymbol{\phi}_{0(intercept)}$
0.36326	-0.17097	0.196487	-0.11968	-0.92704	-0.34195	0.03484

σ^2 estimated as 25.13: log likelihood = -477.89, aic = 985.78

Model:

$$X_t = 0.03484 - 0.1543X_{t-1} + 0.256904X_{t-2} - 0.2199X_{t-3} - 0.23248X_{t-4} + 0.411716X_{t-5} - 0.24344X_{t-6} - 0.1027X_{t-7} + \in_t + 0.36326 \in_{t-1} - 0.17097 \in_{t-2} + 0.196487 \in_{t-3} - 0.11968 \in_{t-4} - 0.92704 \in_{t-5} - 0.34195 \in_{t-6}$$

Checking of Covariance stationarity

We can observe that, the coefficients of MA polynomials are finite. So, we only have to ensure that the AR polynomial is covariance stationary to show that the de-trended time series data is covariance stationary that is we have to check if the roots of the AR polynomial lie outside the unit circle.

0.3349167+1.157047i	-0.8971509+0.510636i	-0.8971509-0.510636i	1.1769546-0.603393i
1.1769546+0.603393i	0.3349167-1.157047i	-3.5993225	

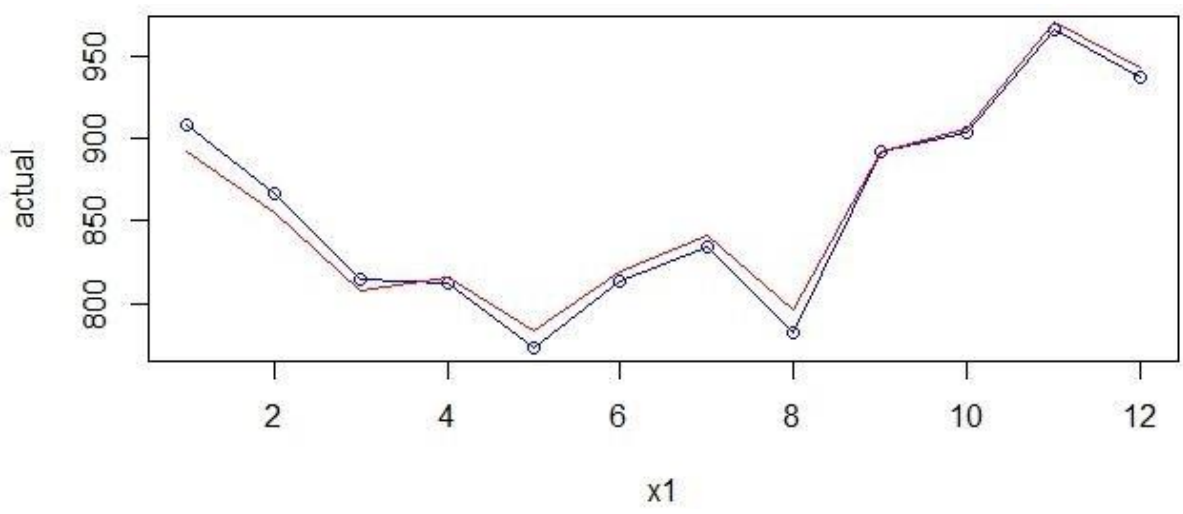
Modulus of all the roots lie outside the unit circle

Forecasting

Forecasting is predicting future values by studying the past patterns of the data. Mostly, we assume that the model is exactly known, including the specific values for all parameters. Although this is never true in practice.

Suppose that the stationary time series model that is fitted to the data $\{x_1, x_2, \ldots, x_n\}$ is known and we would like to predict the future values of the series $X_{n+1}, X_{n+2}, \ldots, X_{n+h}$ based on the realization of the time series up to time n, where , n is the origin of the forecast and h is the lead time.

	Predicted	Actual
[1,]	891.8809	908
[2,]	855.0089	867
[3,]	808.0045	815
[4,]	815.7735	812
[5,]	782.7796	773
[6,]	819.5806	813
[7,]	841.1281	834
[8,]	796.0996	782
[9,]	892.4063	892
[10,]	905.4514	903
[11,]	970.5497	966
[12,]	943.0789	937



Blue-actual Red-predicted

Goodness of fit for the model

The **goodness of fit** of a statistical model describes how well it fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

O_i = an observed values
E_i = an expected values

Test statistics follow chi square with (n-1) degree of freedom .
If the Calculated is greater than tabulated we accept that our model fits the underlying true model else not.
Our values:
Calculated statistic =1.089415
Tabulated statistic=19.675 at 5% level of significance
Hence we conclude that our model is adequate predicting the underlying true model.

Bibliography

1. Data source: https://datamarket.com/data/set/22ox/monthly-milk-production-pounds-per-cow-jan-62-dec-75#!ds=22ox&display=line
2. Introduction to Time Series and Forecasting – P. J. Brockwell and R. A. Davis