Contents

Time Series Forecasting

- Time series modeling deals with the time based data. Time can be years, days, hours, minutes, etc.

- Time series forecasting involves fitting a model on time based data and using it to predict future observations.

- Time series forecasting serves two purposes: understanding the pattern/trend in the time series data and forecasting/extrapolating the future values of it. The **forecast** package in R contains functions which serve these purposes.

- In time series forecasting, the AutoRegressive Integrated Moving Average (ARIMA) model is fitted to the time series data either to better understand the data or to predict future points in the series.

- Components of a time series are level, trend, seasonal, cyclical and noise/irregular (random) variations.

Regression Analysis

Time Series Forecasting

- Figure 1 shows the forecast of 4 future values of 'AirPassengers' data using ARIMA model (available in **forecast** package).

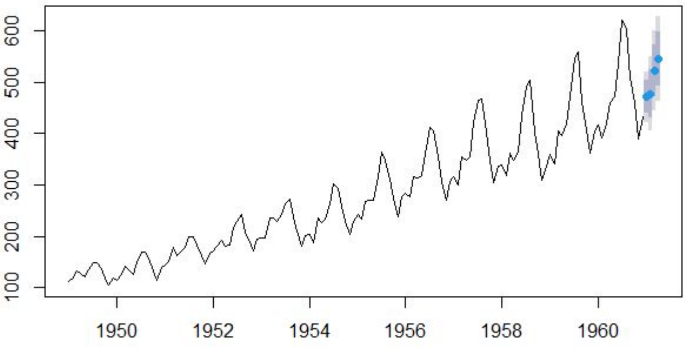

Figure: Forecast from ARIMA(3,1,3) - 'AirPassengers' data

Regression Analysis

Autocorrelation

- As correlation measures the linear relationship between two variables, autocorrelation measures the linear relationship between lagged values of a time series data/variable. The term 'lag' refers to 'time dealy'.
- Figure 2 shows the autocorrelation plot of 'AirPassengers' data obtained using **Acf()** function (available in **forecast** package).

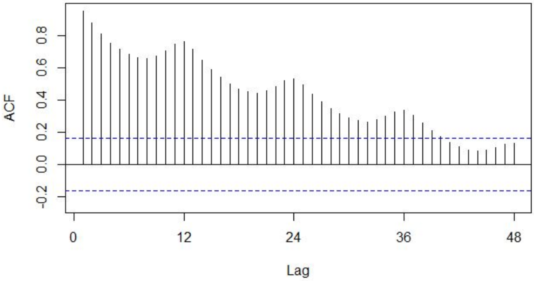

Figure: ACF plot - 'AirPassengers' data

ANOVA - Analysis of Variance

- Analysis of Variance (ANOVA) is a statistical technique for comparing the means of more than 2 sample groups and deciding whether they are drawn from the same population or not.
- The hypothesis is stated as follows:

$$\begin{aligned} \mathbb{H}_0 :&\quad \mu_1 = \mu_2 = \mu_3 = \ldots \\ \mathbb{H}_a :&\quad \mu_1 \neq \mu_2 \neq \mu_3 \neq \ldots \end{aligned}$$

- ANOVA also allows comparision of more than 2 population.
- Assumptions made:
  (i) Samples are independent and randomly drawn from respective populations,
  (ii) Populations are normally distributed, and
  (iii) Variances of the population are equal.

Regression Analysis

ANOVA - Analysis of Variance

- Let X denote the data matrix consisting of samples from r groups such that each column corresponds to one group, $\bar{X}$ denote the mean of all the entries in X, $\bar{x}_j$ denote the mean of all entries in column-j and n_j denote the number of samples in column-j.
- To establish comparison between groups, three variances are considered. They are Sum-of-Squares-Total (SST), Sum-of-Squares-TReatments ($SSTR$) and Sum-of-Squares-Error (SSE):

$$SST = \sum_j \sum_i (X_{i,j} - \bar{X})^2$$

$$SSTR = \sum_j n_j \left(\bar{x}_j - \bar{X}\right)^2$$

$$SSE = \sum_j \sum_i (X_{i,j} - \bar{x}_j)^2.$$

ANOVA - Analysis of Variance

- SST gives the overall variance in the data, $SSTR$ gives the part of the variation within the data due to differences among the groups, and SSE gives the part of the variation within the data due to error. Note that $SST = SSTR + SSE$.
- The ANOVA F-statistic is defined as

  $$F = \frac{MSTR}{MSE}$$

  where $MSTR = SSTR/d.o.f = SSTR/(r-1)$ and $MSE = SSE/d.o.f = SSE/(n-r)$. Note that $n = \sum_j n_j$ is the total number of samples.
- If F-statistic is greater than the critical value, then the null hypothesis is rejected. The critical value is obtained from the F-distribution table using parameters such as significance level (α) and degrees of freedom (d.o.f) of SSTR and SSE.

Question 1.7

Assume there are 3 canteens in a college and the sale of an item in those canteens during first week of February-2021 is as follows:

Table: Data for Question 1.6

Canteen A	Canteen B	Canteen C
40	30	50
60	30	60
70	10	30
30	70	20
50	60	20

Is there a significant difference between the mean sales of the item, at $\alpha = 0.05$?