PROJECT: TIME SERIES ANALYSIS ON MONTHLY CLOSINGS OF THE DOW -JONES INDUSTRIAL INDEX FROM 08-1968 TO 10-1992

NAME - ROCHITA DAS

ROLL NO - 03

Data Description:

We consider monthly closings of the Dow -Jones industrial index from 08-1968 to 10-1992.

The **Dow Jones Industrial Average** also called **DJIA**, the **Industrial Average**, the **Dow Jones**, the **Dow Jones Industrial**, **DJI**, the **Dow 30**, or simply the **Dow**, is a *stock market index* stock market index, and one of several indices created by *Wall Street Journal* editor and Dow Jones & Company co-founder Charles Dow. It is an index that shows how 30 large publicly owned companies based in the United States have traded during a standard trading session in the stock market.

Objective:

We want to analyze how much variation is present in our data set, and want to model it. We also want to forecast .

Summary of data:

Using "summary" measure in "R"-software, we get data summary as follows –

```
  Min.    1st Qu.    Median    Mean    3rd Qu.   Max.

 3537     3673       3753       3755     3850    3978
```

Some graphical representation of closing price is shown as –

To find whether trend is present or not:

- ***Graphically:***

From graph it seems that ***trend*** is present in data.

- ***Test:***

For confirmation we perform **Mann-Kendall test**. The test is as follows:

The purpose of the Mann-Kendall is to statistically assess if there is a monotonic upward or downward trend of the variable of interest over time. A monotonic upward (downward) trend means that the variable consistently increases (decreases) through time, but the trend may or may not be linear.

H0: No monotonic trend vs H1: Monotonic trend is present

Performing the test on our data we get tau = 0.386, ***p value =< 2.22e-16***

Hence, at 1% level of significance the test is accepted, i.e. the sample have *significant trend.*

❖ *Trend estimation :*

From data it seems that upto Oct' of 1980 (i.e. time pt. 147) the data has some quadratic trend pattern and after this time pt. a linear trend can explain trend behavior of the rest. So here we fit spline regression, quadratic equation for treand upto Oct' of 1980 and thereafter linear equation.

Spline 1:

Coefficients:	*(Intercept)*	*time*	$time^2$
	3521.01840	*4.21947*	*-0.01219*

Spline 2 :

Coefficients:	*(Intercept)*	*time*
	3389.644	*1.712*

Original observation with fitted trend value is plotted in the graph.

To find whether seasonality is present or not:

➢ *Graphically:*

We draw **Box-plot** on de-trended data for 12 months. And from plot we see there is no significant seasonality.

➢ *Test:*

For confirmation we perform **Kruskal-Wallis rank sum test**. The test is as follows:

Kruskal-Wallis test is a non-parametric test used for comparing samples from two or more groups. This test does not make assumptions about normality. However, it assumes that the observations in each group come from populations with the same shape of distribution.

H0 : all months have the same mean, i.e. No seasonality is present. Vs. H1 : not H0

Performing the test on de-trend data we get ,

Kruskal-Wallis chi-squared = 4.3212, ***p-value = 0.9596***

Hence, at 1% level of significance the test is accepted, i.e. the sample has not any seasonality.

Hence we can conclude that the de-trended data i.e. entirely consists of random part. So, we can fit model on this non-stationary component.

Model fitting:

➢ *ACF plot:*

In statistics, the **autocorrelation** of a random process describes the correlation between values of the process at different times, as a function of the two times or of the time lag. Let Xi is the value the process at time i. Suppose the process has defined values for mean μi and variance σi2 for all times i. Then the definition of the autocorrelation between times s and t is

$$R(s,t) = \frac{\mathrm{E}[(X_t - \mu_t)(X_s - \mu_s)]}{\sigma_t \sigma_s},$$

where "E" is the expected value operator. If the function R is well-defined, its value must lie in the range [−1, 1], with 1 indicating perfect correlation and −1 indicating perfect anti-correlation

The lag after which, (if) ACF dies out, that lag value is taken as the order of Moving Average (MA) Process to model the data.

From our ACF plot, no prominent decaying pattern is observed. Hence we do not use MA process.

➢ *PACF plot:*

In time series analysis, the **partial autocorrelation** function (PACF) gives the partial correlation of a time series with its own lagged values, controlling for the values of the time series at all shorter lags.

Given a time series z_t, the partial autocorrelation of lag k, denoted $\alpha(k)$, is the autocorrelation between $z_t$and $z_{t+k}$with the linear dependence of $z_t$on z_{t+1} through z_{t+k-1} removed.

$$\alpha(1) = \operatorname{Cor}(z_{t+1}, z_t),$$

$$\alpha(k) = \operatorname{Cor}(z_{t+k} - P_{t,k}(z_{t+k}),\ z_t - P_{t,k}(z_t)),\ \text{for } k \geq 2,$$

where $P_{t,k}(x)$denotes the projection of $x$onto the space spanned by $x_{t+1}, \ldots, x_{t+k-1}$

The point on the plot where the partial autocorrelations for all higher lags are essentially zero is the order of Auto Regressive (AR) model.

From our PACF plot, no prominent decaying pattern is observed. Hence we do not use AR process.

• *Graphically:*

We plot **Normal Q-Q Plot**. Here, sample quantile is plotted against quantile of Normal distribution. From the plot, it is clear that data does not lie in a straight line. And data has heavy tail observations. Hence, the data plot says to proceed with **ARCH or GARCH** model.

• *Test of Normality:*

We further test, whether the data comes from Normal Distribution or not by **Kolmogorov-Smirnov Test.**

The Kolmogorov–Smirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution.

H0: the sample comes from reference distribution vs. H1: not H0

Taking reference as Normal Distribution, wrt our data the test is as follows,

D = 0.543, ***p-value < 2.2e-16***

Hence, at 1% level of significance the test is rejected, i.e. the sample does not come from Normal Distribution.

Moreover,

- ✓ The plot shows, past volatility affects the present volatility, large fluctuation is followed by small one and again big one and such fluctuations are continuous. This is the evidence of **Volatility**.
- ✓ If expected return is assumed to follow a standard regression or time series model, the variance is immediately considered to be constant over time, which might not be the reality. This is **Conditional Heteroscedasticity.**
- ✓ Moreover the use of an exogenous variable to explain changes in variance is usually not appropriate. This is **Structural Misspecification**.
- ✓ The plot gives heavy tailed distribution also large and small errors tend to cluster together. This is **Temporal Clustering**.

Due to these reasons we opt for **ARCH or GARCH** model to fit the data.

GARCH MODEL :

In that case, the **GARCH (p, q)** model (where p is the order of the GARCH terms σ^2 and q is the order of the ARCH terms ϵ^2), is given by

$$y_t = x_t'b + \epsilon_t$$

$$\epsilon_t|\psi_t \sim \mathcal{N}(0, \sigma_t^2)$$

$$\sigma_t^2 = \omega + \alpha_1\epsilon_{t-1}^2 + \cdots + \alpha_q\epsilon_{t-q}^2 + \beta_1\sigma_{t-1}^2 + \cdots + \beta_p\sigma_{t-p}^2 = \omega + \sum_{i=1}^{q}\alpha_i\epsilon_{t-i}^2 + \sum_{i=1}^{p}\beta_i\sigma_{t-i}^2$$

To find the appropriate order of GARCH, we put value from 0 to 2 in p, q in different combinations and find the AIC of the fitted model. The model with minimum **AIC** is the best model to fit the data.

Table:

Order	AIC
p =1, q=0	3142.018
p =0, q=1	3015.017
p =1, q=1	3099.372
p =1, q=2	3085.891
p =2, q=1	3085.518
p =2, q=2	3083.724

AIC is minimum for p=0, q=1. Hence **GARCH (0, 1)** i.e. **ARCH(1)** fits the data in the best way.

On fitting ARCH (1) we get,

```
      Estimate      Std. Error    t value    Pr(>|t|)

a0    508.2008      110.8241      4.586      4.53e-06

a1    0.8452        0.2084        4.057      4.97e-05
```

Forecast:

With sample size n= 291, we ***forecast non stationarity*** for (n+1) th value as

```
 meanForecast    meanError    standardDeviation    lowerInterval    upperInterval

 -8.171636e-14   37.76083     37.76083             -74.00987        74.00987
```

From ***trend equation*** (spline 2)

forcasted trend : *3889.548*

Hence we can ***forecast*** for ***(n+1) th*** i.e. 292 th ***observation*** simply adding the forecasted value of non-stationary part and trend part , it becomes : *3889.548*

The forecasted non-stationary part with confidence interval is shown in the graph.

Data Source :

https://datamarket.com/data/set/22v9/monthly-closings-of-the-dow-jones-industrial-index-aug-1968-aug-1981#!ds=22v9&display=line