# Part A: TIME SERIES ANALYSIS

## I. INTRODUCTION TO TIME SERIES ANALYSIS

A time series is a set of numerical data points in sequential orders. Time series analysis indicates that data points taken over time that might have an intrinsic pattern (such as autocorrelation, trend, or seasonal variation) should be taken into account. These data points and associated patterns are used to forecast future events. It may be helpful to see how the commodity, security, or economic component shifts over time. Also, a study of the time series on foreign currency exchange rates will help to assess if the exchange rate is going through peaks and troughs regularly each year and the right time to do the foreign investments. Research in this field will involve the implementation of the observed exchange rates and their association with the chosen season.

## II. STATEMENT OF OBJECTIVES AND DESCRIPTION OF DATA

The purpose of this project is to apply a time series analysis on the monthly average exchange rate of Euro (European Currency Unit) versus Pounds Sterling from 1971 till 2020 to determine the supply and demand of Euro over Pounds in the global currency market.

Entire operations in this project are performed in the RStudio application.

### A. Data Description

Dataset used in this analysis is sourced from https://ec.europa.eu/eurostat. The dataset contains the exchange rate of one unit Euro into Pounds sterling (in Pounds) from January 1971 till November 2020.

### B. Variables

The variable used in this analysis are as follows:

**Index**- Monthly Time from 1971 to 2020.

**CoreData (Dependent Variable)**- Exchange rate of Euro in Pounds.

## III. TIME SERIES ANALYSIS:

To proceed with model building and forecasting through the best-fitted time series model, pre-analysis of the data is done to look for various insights from it.

### A. Plotting Time Series Data:

Using R, Normal Data is converted into time-series data by placing it into a time-series object as shown in the below screen capture. Now, it is plotted to identify the underlying pattern in the data as shown in Figure 1 below.

```
> exchange_rate <- read_excel("C:/Users/Ankit/Desktop/Data Analytics/Statistics/TABA/Ankit_Time_series.xlsx")
> str(exchange_rate)
tibble [599 x 1] (S3: tbl_df/tbl/data.frame)
 $ Exchange Rate of Euros to Pound (in Pounds): num [1:599] 0.427 0.425 0.425 0.426 0.428 ...
> exchange_rate <- ts(exchange_rate, start=c(1971,1), frequency=12)
> head(exchange_rate)
          Jan     Feb     Mar     Apr     May     Jun
1971 0.42704 0.42549 0.42550 0.42551 0.42802 0.42846
> start(exchange_rate)
[1] 1971    1
> end(exchange_rate)
[1] 2020   11
> frequency(exchange_rate)
[1] 12
> plot.ts(exchange_rate,main="Monthly Exchange Rate of Euro to Pound from 1970 to 2020 ")
```

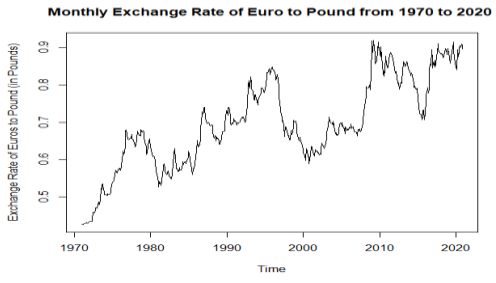



**Figure 1: Time-series Plot**

The above graph (Figure 1), shows an increasing trend pattern i.e. supply and demand of the Euro in the global currency market is increasing with years against pounds. Various rise and fall that are not of a fixed frequency show the existence of the cyclical pattern in the data.

### B. Seasonal Plots

To check for seasonality, data points are plotted against the individual seasons in which they were observed. Below the graph (Figure 2 & Figure 3) means that each month is relatively similar and reveals no apparent trend and doesn't offer any evidence of seasonality.

```
> ggsubseriesplot(exchange_rate)+ ylab("Exchange Rate of Euros in £ ") + ggtitle("Seasonal Subseries Plot: Exchange Rate of Euro to Pounds sterling from 1971 to 2020")
```

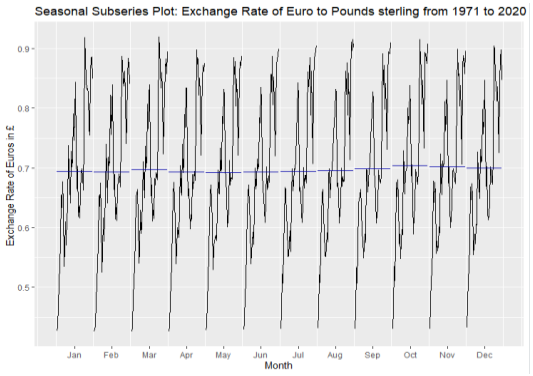



**Figure 2:Seasonal Sub-Series Plot**

```
> ggseasonplot(exchange_rate,year.labels = TRUE,year.labels.left = TRUE)+ ylab("Exchange Rate of Euros in £ ") + ggtitle("Seasonal Plot: Exchange Rate of Euro to Pounds sterling from 1971 to 2020")
```

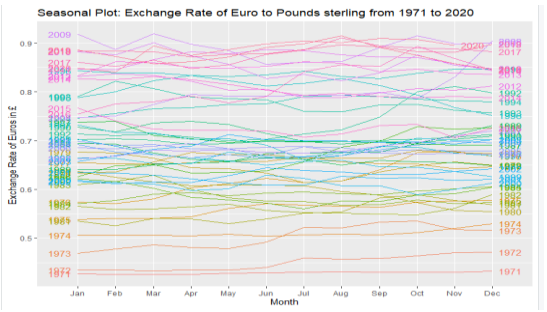



**Figure 3:Seasonal Plot**

### C. Seasonal Decomposition

Time series decomposition means thinking of a series as a mixture of observed, trend, seasonal and random components. Decomposition provides a valuable abstract paradigm for reasoning about time series in general and for a deeper comprehension of problems during time-series research and forecasting.

### Multiplicative Model:

A multiplicative model is non-linear, such as quadratic or exponential. This model is suitable in this situation as the variations depend on the stage of the time series. Changes can rise or decrease over time. Even, after adding both the Multiplicative and the Additive Models to the efficiency measure, the multiplicative decomposition is doing well in this situation.

$$Y_t = Trend_t * Seasonal_t * Irregular_t$$

```
> fit.decmult<-decompose(exchange_rate,type = "multiplicative")
> plot(fit.decmult)
```

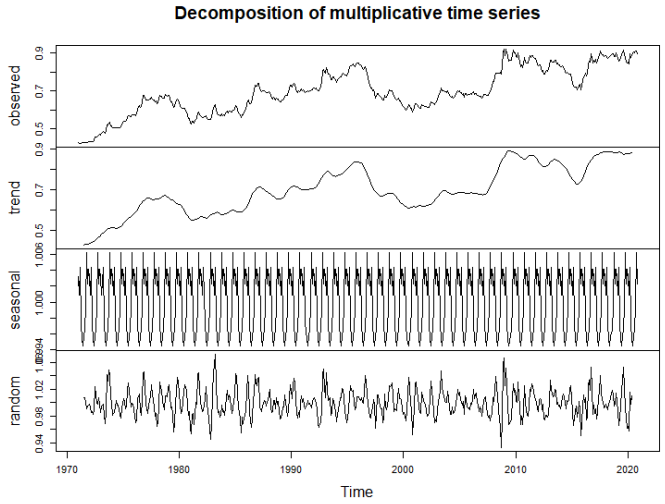



**Figure 4:Multiplicative Decomposition Plot**

## IV. MODEL BUILDING AND DIAGNOSTICS:

After the time-series analysis, the models will be built based on the above analysis, and according to their accuracies, the best-fit model is chosen and the exchange rate of Euros to Pounds is predicted for the next few periods.

### A. Model 1- Holt-Winters Seasonal Multiplicative Method

Holt-Winters is a way of modeling three elements of the time series: a typical value (average), a slope (trend) over time, and a cyclical pattern of repetition (seasonality), with the corresponding smoothing parameters α, β, and γ. Holt-Winters uses exponential smoothing to encode a lot of values from the past and use them to estimate "typical" values for the current and the future. The data shows a trend and cyclical pattern, hence applying the model to forecast the exchange rate.

```
> #holtwinters Model Multiplicative
> exchangefit_multiplicative<-hw(exchange_rate,seasonal = "multiplicative")
> autoplot(exchange_rate)+autolayer(exchangefit_multiplicative,series="HW multiplicative forecasts",PI=FALSE)+guides(color=guide_leg
end(title="Forecast"))
> summary(exchangefit_multiplicative)

Forecast method: Holt-Winters' multiplicative method

Model Information:
Holt-Winters' multiplicative method

Call:
 hw(y = exchange_rate, seasonal = "multiplicative")

  Smoothing parameters:
    alpha = 0.8247
    beta  = 1e-04
    gamma = 0.1747

  Initial states:
    l = 0.4257
    b = 7e-04
    s = 0.9933 1.0434 1.015 0.9793 0.9908 0.9638
           1.0088 0.9874 0.979 1.0249 0.9841 1.0301

  sigma:  0.0205

     AIC     AICc      BIC
-1258.998 -1257.944 -1184.278

Error measures:
                      ME        RMSE         MAE          MPE     MAPE      MASE      ACF1
Training set 5.356371e-05 0.01396242 0.01046389 -0.001662091 1.515406 0.2483196 0.3094069

Forecasts:
         Point Forecast     Lo 80     Hi 80     Lo 95     Hi 95
Dec 2020      0.9028297 0.8790820 0.9265775 0.8665107 0.9391488
Jan 2021      0.9109186 0.8798671 0.9419701 0.8634294 0.9584078
Feb 2021      0.9106160 0.8738330 0.9473989 0.8543613 0.9668707
Mar 2021      0.9242204 0.8818593 0.9665815 0.8594347 0.9890060
Apr 2021      0.9122734 0.8660303 0.9585165 0.8415507 0.9829961
```

**Figure 5: Holt-Winters Multiplicative Plot**

The above graph (Figure 5), displays the Holt-Winters Multiplicative Model forecast and the code snippet shows the forecast accuracy i.e. RMSE= 0.01396242 and MAPE (Mean avg. percentage error) = 1.515406. The RMSE of both the additive and the multiplicative versions are compared. In this case, the multiplicative seasonality method is ideally suited to the findings. The explanation for this is that the time plot indicates that trend variance in data rises as the level of the sequence increases.

α = 0.8247
β = 1e-04
γ = 0.1747

### B. Model 2- Holt's Linear Trend:

Holt's two-parameter model, also known as linear exponential smoothing, is a common smoothing model for trend data forecasting. Holt's model has three different equations that function together to produce a final prediction. This model is ideal for a series in which there are a linear trend and no seasonality. Its smoothing parameters are the level and the pattern, which are not limited by the values of each other. It is ideal to apply this model in this time-series data as it is a linear trend and doesn't show any seasonality in it. Below are the forecasts, summary, and the forecast plot of this model.

```
> #Holt
> holt_exchange_Rate<-holt(exchange_rate)
> summary(holt_Rate)

Forecast method: Holt's method

Model Information:
Holt's method

Call:
 holt(y = exchange_rate)

  Smoothing parameters:
    alpha = 0.9986
    beta  = 1e-04

  Initial states:
    l = 0.4003
    b = 9e-04

  sigma:  0.0127

     AIC     AICc      BIC
-1397.281 -1397.180 -1375.305

Error measures:
                      ME        RMSE         MAE          MPE     MAPE      MASE      ACF1
Training set -0.0001082729 0.01262189 0.009152879 -0.01968258 1.29496 0.2172079 0.2259475

Forecasts:
         Point Forecast     Lo 80     Hi 80     Lo 95     Hi 95
Dec 2020      0.8969952 0.8807653 0.9132250 0.8721737 0.9218166
Jan 2021      0.8979230 0.8749855 0.9208604 0.8628431 0.9330028
Feb 2021      0.8988508 0.8707634 0.9269381 0.8558949 0.9418066
Mar 2021      0.8997786 0.8673483 0.9322088 0.8501808 0.9493764
Apr 2021      0.9007064 0.8644490 0.9369638 0.8452555 0.9561573
May 2021      0.9016342 0.8619161 0.9413523 0.8408906 0.9623778
Jun 2021      0.9025620 0.8596608 0.9454631 0.8369504 0.9681736
Jul 2021      0.9034898 0.8576254 0.9493542 0.8333463 0.9736333
Aug 2021      0.9044176 0.8557696 0.9530656 0.8300169 0.9788183
Sep 2021      0.9053454 0.8540641 0.9566267 0.8269175 0.9837733
> plot(holt_Rate)
```

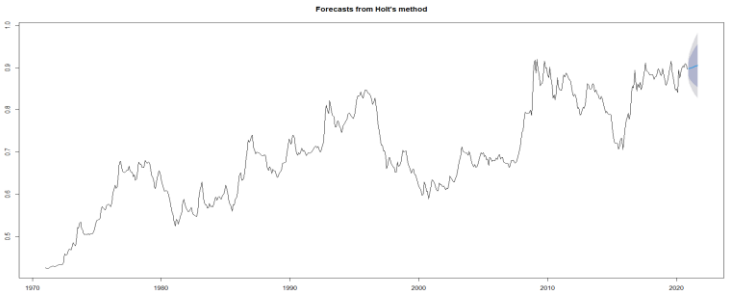



**Figure 6: Holt's Linear Plot**

The above graph (Figure 6), displays the Holt's Linear Model forecast and the code snippet shows the forecast accuracy i.e. RMSE= 0.01262189 and MAPE (Mean avg. percentage error) = 1.29496.

α = 0.9986 , exponential decay for the level
β = 1e-04 , exponential decay for the slope

### C. Model 3 - Random Walk Model (Naïve Model)

Random walk models are commonly used for non-stationary data, especially financial and economic data. It generally has long stretches of apparent up-and down-trends and abrupt and unforeseeable shifts of course. This is why Forecasts from a random walk model are equal to the last observation. This is also called a naive forecast which can be achieved using the 'naive()' function. Below are the summary and the plot of the naïve forecast.

$$y_t = y_{t-1} + E_T$$

```
> #Naive
> naive_exchange<-naive(exchange_rate,h=4)
> summary(naive_exchange)

Forecast method: Naive method

Model Information:
Call: naive(y = exchange_rate, h = 4)

Residual sd: 0.0126

Error measures:
                     ME       RMSE        MAE       MPE     MAPE      MASE      ACF1
Training set 0.0007842977 0.01260726 0.009064866 0.1087929 1.278111 0.2151193 0.2269018

Forecasts:
         Point Forecast     Lo 80     Hi 80     Lo 95     Hi 95
Dec 2020        0.89605 0.8798931 0.9122069 0.8713402 0.9207598
Jan 2021        0.89605 0.8732008 0.9188992 0.8611051 0.9309949
Feb 2021        0.89605 0.8680655 0.9240345 0.8532514 0.9388486
Mar 2021        0.89605 0.8637363 0.9283637 0.8466304 0.9454696
> plot(naive_exchange)
```

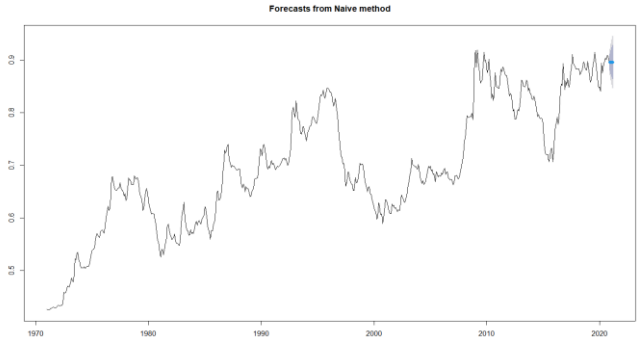



**Figure 7: Naïve Forecast Plot**

The above graph (Figure 7), displays the naïve forecast and the code snippet shows the forecast accuracy i.e. RMSE-0.01260726 and MAPE (Mean avg. percentage error) - 1.278111. It shows a slight improvement from Holt's Linear Trend Model.

### D. Model 4 – ARIMA MODEL (OPTIMUM MODEL)

ARIMA stands for 'Auto-Regressive Integrated Moving Average' is a class of models that 'explain' a given time series based on its historical values, i.e. its lags and lagged forecast errors, such that the equation can be used to estimate future values.

ARIMA models are designed to fit stationary time-series or the ones that can be made stationary. The lags of the stationary series of the forecast equation are called "auto-regressive" terms, the lags of the forecast errors are called "moving average" terms, and the time series that has to be differenced to be stationary is considered to be the "integrated" version of the stationary series.

Any 'non-seasonal' time series that shows trends and is not spontaneous white noise can be modeled with ARIMA models. If a time series has seasonal variations, SARIMA, short for 'Seasonal ARIMA' is applied.

The ARIMA model is characterized by three terms: p, d, q.

where,
p is the order of the AR term
q is the order of the MA term
d is the number of differences required to make the time series stationary

To use the above model, values of p,d,q are determined below and performed certain tests on the time-series dataset and made necessary changes wherever required to make the time-series data ready for the ARIMA model:

#### i. Stationary Testing:

Since the ARIMA model works on stationary data, checking for the stationarity of the used time-series data in this project. A time series is stationary if its properties, such as mean and variance, do not change over time. As per our time-series analysis plot (Figure 1), it is showing a trend with the existence of a cyclic pattern, which is not stationary.

Below is the test to check for stationarity:

ADF (Augmented Dickey-Fuller test):

$H_0$: Time-series data is NOT stationary

$H_1$: Time-series data is stationary.

As per the below test, P-value is not significant, making the time-series data not stationary.

```
> #Assess stationarity
> adf.test(exchange_rate)

        Augmented Dickey-Fuller Test

data:  exchange_rate
Dickey-Fuller = -3.0504, Lag order = 8, p-value = 0.1337
alternative hypothesis: stationary
```

For making a non-stationary time series stationary, differences between consecutive observations are computed. This is known as differencing. It helps in stabilizing the mean of a time series by removing changes in the level of a time-series, as a result eliminating trend and seasonality.

Checked for the order of differencing:

The order of differencing in the applied time-series data is 1.

```
> #Check the order of differencing required
> ndiffs(exchange_rate)
[1] 1
```

Applied Differencing:

After applying differencing to the time-series data, the data became stationary and ready for the ARIMA model, also shown by the significant p-value, as a significant p-value is the proof of stationarity. Plotted the differenced time-series as shown below in Figure 8.

```
> #Apply differencing and Plot the differenced  Time Series
> dRate <- diff(exchange_rate)
> plot(dRate)
> #Assess stationarity of the differenced Time-Series
> adf.test(dRate)

        Augmented Dickey-Fuller Test

data:  dRate
Dickey-Fuller = -7.4772, Lag order = 8, p-value = 0.01
alternative hypothesis: stationary
```

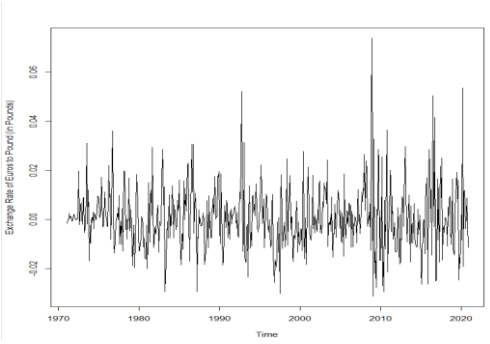


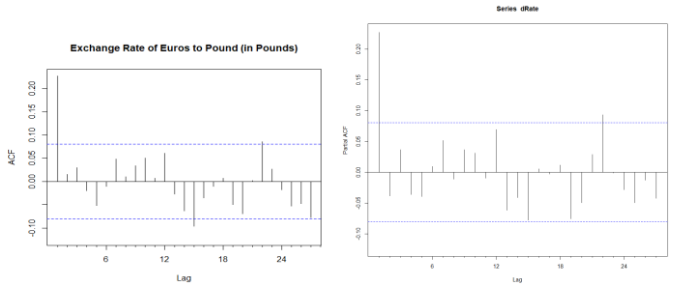

**Figure 8: Stationary Plot**

The differenced time series is called an Integrated series. With this value of d becomes 1, as the order of differencing is 1 as shown in the above code snippet.

#### ii. Auto-Correlation Function and Partial Auto-Correlation Function Test:

After rendering the dataset stationary, we switch to the second stage of the process: p, d, q for the ARIMA model. This can be achieved by looking at the PACF and ACF graphs.

```
> #ACF/PACF plots. Choosing p and q
> Acf(dRate)
> Pacf(dRate)
```




Since there is a single spike at lag 1 in both the ACF and PACF plots, setting p=1 and q=0 as it is positive (this is an AR(1) signature).

### iii. Fitting ARIMA Model:

With values p=1, d=1, and q=0, fitting the ARIMA model. The number of significant lags in the ACF and PACF plots is converted into the corresponding p&q.

```
> #Fitting an ARIMA model
> exchange_arima <- arima(exchange_rate, order=c(1,1,0))
> summary(exchange_arima)

Call:
arima(x = exchange_rate, order = c(1, 1, 0))

Coefficients:
          ar1
      0.2299
s.e.  0.0398

sigma^2 estimated as 0.0001505:  log likelihood = 1783.06,  aic = -3562.13

Training set error measures:
                      ME       RMSE        MAE       MPE     MAPE      MASE        ACF1
Training set 0.0005993689 0.01225854 0.008830642 0.0851166 1.243331 0.9741613 0.006176318
```

The above code snippet displays the summary of the ARIMA forecast and the forecast accuracy i.e. RMSE= 0.01224673 and MAPE (Mean avg. percentage error) = 1.225854. It has AIC=-3562.13. It shows the minimal RMSE and MAPE after comparing all the four models above. ARIMA model outperforms all the above applied models.

### iv. Forecast With ARIMA:

Figure 9 below displays the ARIMA forecast and the code snippet displays the forecasted values for the next 4 months.

```
> #Forecasting with the fitted model
> forecast(exchange_arima, 4)
         Point Forecast     Lo 80     Hi 80     Lo 95     Hi 95
Dec 2020      0.8934379 0.8777148 0.9091609 0.8693915 0.9174842
Jan 2021      0.8928372 0.8679136 0.9177609 0.8547198 0.9309547
Feb 2021      0.8926991 0.8606365 0.9247617 0.8436636 0.9417346
Mar 2021      0.8926674 0.8546861 0.9306486 0.8345800 0.9507547
> plot(forecast(exchange_arima, 4), xlab="Year", ylab="Exchange Rate")
```

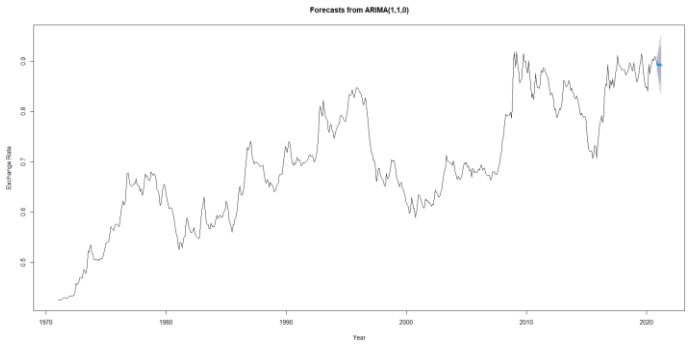
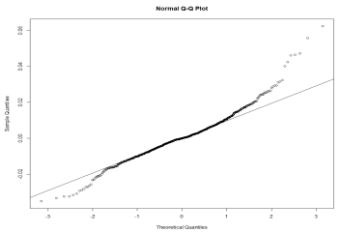



**Figure 9: ARIMA Forecast Plot**

### v. Evaluating Model Fit:

**Normal Q-Q Plot**

As per the Normal Q-Q plot, the residuals are normally and independently distributed and have no relationship between them.




**Ljung-Box Test:**

P-value is not significant, which shows autocorrelations don't differ from zero and the ARIMA model has fit the data well.

```
> Box.test(exchange_arima$residuals, type="Ljung-Box")

        Box-Ljung test

data:  exchange_arima$residuals
X-squared = 0.022965, df = 1, p-value = 0.8795
```

**Check Residual Method:**

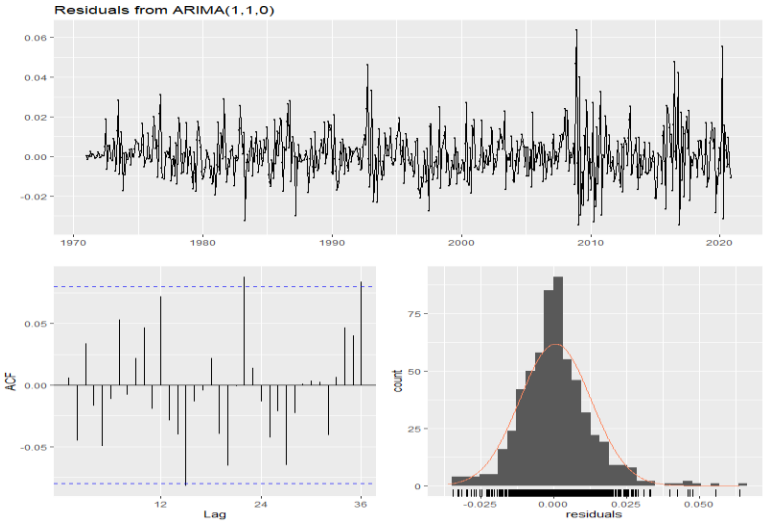
The residual autocorrelations don't differ significantly from 0 and we do not see any significant spikes in the first few lags.

```
> checkresiduals(exchange_arima)

        Ljung-Box test

data:  Residuals from ARIMA(1,1,0)
Q* = 25.243, df = 23, p-value = 0.3379

Model df: 1.   Total lags used: 24
```




## V. Conclusion:

This project focused on applying Time-Series analysis to forecast the exchange rate of 1 unit Euro in Pound sterling (in Pounds). ARIMA model outperformed all the four models applied. It is observed that the currency value of the Euro has an increasing trend in years against pounds sterling, making it more supplied and demanded in the global currency market than pounds.

## VI. References

[1] Ayekple, Yao & Harris, Emmanuel & Frempong, Nana & Amevialor, Joshua. (2015). Time Series Analysis of the Exchange Rate of the Ghanaian Cedi to the American Dollar. Journal of Mathematics Research. 7. 10.5539/jmr.v7n3p46.

# Part B – LOGISTIC REGRESSION ANALYSIS

## I. INTRODUCTION TO LOGISTIC REGRESSION

Logistic regression is a classification algorithm. They work on a qualitative response variable, unlike linear regression which works on quantitative response variables. It is used to predict a categorical output based on a series of independent variables which can be qualitative and quantitative as well. If the response variable is dichotomous, binary logistic Regression is performed, else multinomial logistic regression is performed. The approach used for classification first estimates the probability of each of the categories of the qualitative variable as the basis for classification. In this way, they function the same as regression methods.

Logistic Regression Equation:

$$E(Y)=\exp(\beta_0+\beta_1X_1+...+\beta_kX_k)/1+\exp(\beta_0+\beta_1X_1+...+\beta_kX_k)$$

E(y) as a Probability in Logistic Regression:

If the two values of y are 0 or 1, the value of E(y) provides the probability that y = 1 given a particular set of values for $X_1, X_2, \ldots, X_K$.

## II. STATEMENT OF OBJECTIVES AND DESCRIPTION OF VARIABLES

As of October 2020, 59% of the global population were active Internet users. Mobile has also become the world's biggest Internet access channel, with mobile Internet subscribers responsible for 91% of total Internet users. By Using Logistic Regression, The objective of this project is to analyze Internet Usage based on multiple independent variables such as Owning a Mobile Phone, Standard Of Living, Access to Desktop/Computer in the household or at Work, Age, Read Some English, Highest Level of School ( Below or Above Graduation), and Urbanity.

Entire operations in this project are performed in the IBM SPSS Statistics application.

### i) Data Description

The Dataset used in this project is prepared and sourced from https://www.pewresearch.org/download-datasets/. The Independent variable "Highest level of education" initially was having more than two categories, which was later merged into two categories named "Below HighSchool" and "Above HighSchool". The dataset contains 8 columns having 7 independent and one dependent variable.

### ii) Description of Variables

The variables used in this regression are as follows:

A. Dependent Variable:

**Type-** Categorical Variable:

**Variable Name-** Do you currently use the internet — yes or no?

Categories:

1) Yes (1)
2) No (2)

B. Independent Variable:

**Type-** Categorical Variables:

**Variable Name-** 1. Do you own a mobile phone — yes or no?

Categories:

1) Yes (1)
2) No (2)

2. Compared to your parents when they were the age you are now, do you think your own standard of living now is better, worse, or about the same as theirs was?

Categories:

1) Better (1)
2) Worse (2)

3. Do you have access to a working desktop computer, laptop or tablet in your household or at work?

Categories:

3) Yes, at home (1)
4) No, neither at home nor at work (2)

4. Can you read at least some English — yes or no?

Categories:

1) Yes (1)
2) No (2)

5. What is the highest level of school you have completed

Categories:

1) Below HighSchool (1)
2) Above HighSchool (2)

6. Urbanity

Categories:

1) Urban (1)
2) Suburban (2)

**Type-** Continuous Variables:

**Variable Name-** 1. How old were you at your last birthday? Referred to as Age Variable in this report.

## III. ASSUMPTIONS UNDERTAKEN

To perform Logistic regression, there are certain which needs to be met. Below are the assumptions that were undertaken to perform the regression analysis.

- **Assumption 1: Outcomes of Dependent Variable are Mutual Exclusive:**

Dependent Variables should fit into one of the two distinct categories. Mutual Exclusivity can be understood as the existence of one category indicates the absence of another category. Dependent Variable consists of two values i.e. Yes and No. No observations should have both the values of the dependent variable.

As per the classification table shown below, there are 2302 observations, of which 987 people reported "No" and 1315 reported "Yes" on currently using the internet. There are no existing observations with both categories reported. Hence, this assumption is fulfilled and logistic regression can be performed by checking other assumptions.

**Dependent Variable Encoding**

| Original Value | Internal Value |
|---|---|
| Yes | 0 |
| No | 1 |

**Classification Table[a,b]**

| | | | Predicted: Do you currently use the internet — yes or no? | | |
|---|---|---|---|---|---|
| | Observed | | Yes | No | Percentage Correct |
| Step 0 | Do you currently use the internet — yes or no? | Yes | 0 | 987 | .0 |
| | | No | 0 | 1315 | 100.0 |
| | Overall Percentage | | | | 57.1 |

a. Constant is included in the model.

b. The cut value is .500

- **Assumption 2: Sample Size:**

The larger the sample size, the more accurate the results of a study can be assumed to be. According to this assumption, the sample size should be large enough to perform logistic regression. Limited samples with a large number of predictors can be troublesome. It is considered best when 20 cases per predictor are available to perform the regression. As there are 2302 observations present in the undergoing analysis, it is sufficient to perform the logistic regression and meeting the assumption.

**Case Processing Summary**

| Unweighted Cases[a] | | N | Percent |
|---|---|---|---|
| Selected Cases | Included in Analysis | 2302 | 100.0 |
| | Missing Cases | 0 | .0 |
| | Total | 2302 | 100.0 |
| Unselected Cases | | 0 | .0 |
| Total | | 2302 | 100.0 |

a. If weight is in effect, see classification table for the total number of cases.

- **Assumption 3: To check Multicollinearity:**

It exists when the predictor variables are correlated. Very often, predictor variables are strongly related to the dependent variable but should not be strongly related to each other. The typical multicollinearity approach is to drop one of the strongly correlated predictor variables and recompute the regression equation. Multicollinearity can be checked by the correlation matrix. Also, we can use VIF (Variance Inflation Factor) to look for multicollinearity. Values in the Correlation matrix between -0.70 and +0.70 are considered satisfactory. As per below Correlation matrix, no values are breaching the standard range. The highest value present in the correlation matrix below is .228 which is highly acceptable. Hence, multicollinearity is not present and regression analysis can be performed with all the available predictors.

**Correlation Matrix**

| | | Constant | Do you own a mobile phone — yes or no? | Compared to your parents when they were the age you are now, do you think your own standard of living now is better, worse, or about the same as theirs was? | Do you have access to a working desktop computer, laptop or tablet in your household or at work? | How old were you at your last birthday? | Can you read at least some English — yes or no? | What is the highest level of school you have completed | Urbanity |
|---|---|---|---|---|---|---|---|---|---|
| Step 1 | Constant | 1.000 | | | | | | | |
| | Do you own a mobile phone — yes or no? | -.357 | 1.000 | | | | | | |
| | Compared to your parents when they were the age you are now, do you think your own standard of living now is better, worse, or about the same as theirs was? | -.041 | .010 | 1.000 | - | | | | |
| | Do you have access to a working desktop computer, laptop or tablet in your household or at work? | -.608 | .031 | -.008 | 1.000 | | | | |
| | How old were you at your last birthday? | -.378 | -.013 | .013 | .204 | 1.000 | | | |
| | Can you read at least some English — yes or no? | -.514 | -.025 | -.003 | -.015 | -.035 | 1.000 | | |
| | What is the highest level of school you have completed | -.511 | .114 | .024 | .055 | -.025 | .228 | 1.000 | |
| | Urbanity | -.233 | .033 | .016 | -.004 | .089 | .011 | .060 | 1.000 |

- **Assumption 4: Outliers:**

No big outliers, high leverage points, or extremely potent points should be present. These points are unusual when doing logistic regression analysis, and thus need to be tackled to avoid a negative impact on the regression equation, which is used to forecast dependent variables based on all independent variables. It can be analyzed using the cook's distance, which should be less than 1.

As per the below residual statistics, the minimum and maximum cook's distance is .000 and .023 respectively, which is far less than one. Hence, meeting this assumption to perform the regression.

**Residuals Statistics[a]**

| | Minimum | Maximum | Mean | Std. Deviation | N |
|---|---|---|---|---|---|
| Predicted Value | .96 | 2.40 | 1.57 | .301 | 2302 |
| Std. Predicted Value | -2.031 | 2.742 | .000 | 1.000 | 2302 |
| Standard Error of Predicted Value | .013 | .085 | .020 | .008 | 2302 |
| Adjusted Predicted Value | .96 | 2.40 | 1.57 | .301 | 2302 |
| Residual | -1.220 | 1.040 | .000 | .393 | 2302 |
| Std. Residual | -3.102 | 2.645 | .000 | .999 | 2302 |
| Stud. Residual | -3.107 | 2.649 | .000 | 1.000 | 2302 |
| Deleted Residual | -1.224 | 1.044 | .000 | .394 | 2302 |
| Stud. Deleted Residual | -3.113 | 2.653 | .000 | 1.000 | 2302 |
| Mahal. Distance | 1.369 | 106.485 | 5.997 | 10.035 | 2302 |
| Cook's Distance | .000 | .023 | .000 | .001 | 2302 |
| Centered Leverage Value | .001 | .046 | .003 | .004 | 2302 |

a. Dependent Variable: Q9. Do you currently use the internet — yes or no?

- **Assumption 5: Independent Observations:**

This implies that the residual does not have a sequence. If the residuals are correlated, this is called Autocorrelation and can be checked using the Durbin-Watson statistic. It varies from 1 to 3 and is considered acceptable when it is close to 2. According to the table below, successive residuals are distinct as the observed value is 1.764 which is close to 2.

**Model Summary[b]**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Durbin-Watson |
|---|---|---|---|---|---|
| 1 | .609[a] | .370 | .369 | .393 | 1.764 |

a. Predictors: (Constant), What is the highest level of school you have completed, Compared to your parents when they were the age you are now, do you think your own standard of living now is better, worse, or about the same as theirs was?, Do you have access to a working desktop computer, laptop or tablet in your household or at work?, How old were you at your last birthday?, Can you read at least some English — yes or no?, Do you own a mobile phone — yes or no?

b. Dependent Variable: Do you currently use the internet — yes or no?

## IV. Model Building and Diagnostics:

After meeting all the assumptions, based on the dependent and 7 independent variables models will be built. Models can be built by using the Binary Logistic option from the Regression menu under the Analyze tab.

- **Model 1:**

First Model is built by all 7 independent variables. Block 0, Omnibus Test of Model Coefficients, Model Summary, Hosmer & Lemeshow Test, Classification Table, and Variables in the equation of the model are represented below:

**Block 0: Beginning Block**

This Block represents the model with no independent variables. This means that it shows Internet usage without considering any independent variable as shown below. To see an improvement in classification accuracy, independent variables are added in subsequent blocks. The classification table shows that 987 out of 2302 people said "Yes" and 1315 said "No" to if they use the internet, which gives an accuracy of 57.1%.

**Classification Table[a,b]**

| | | | Predicted: Do you currently use the internet — yes or no? | | |
|---|---|---|---|---|---|
| | Observed | | Yes | No | Percentage Correct |
| Step 0 | Do you currently use the internet — yes or no? | Yes | 0 | 987 | .0 |
| | | No | 0 | 1315 | 100.0 |
| | Overall Percentage | | | | 57.1 |

a. Constant is included in the model.

b. The cut value is .500

**Variables in the Equation**

| | | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| Step 0 | Constant | .287 | .042 | 46.416 | 1 | .000 | 1.332 |

## Block 1:

The Omnibus Testing of Model Coefficients is used to verify that the current model (including explanatory variables) is an improvement over the baseline model. It uses chi-square tests to see whether there is a substantial difference between the Log-likelihoods of the baseline model and the current model.

**Omnibus Tests of Model Coefficients**

| | | Chi-square | df | Sig. |
|---|---|---|---|---|
| Step 1 | Step | 1120.429 | 7 | .000 |
| | Block | 1120.429 | 7 | .000 |
| | Model | 1120.429 | 7 | .000 |

Now, a global hypothesis test is conducted which is tested by a statistic that is distributed approximately $X^2$ with Degrees of Freedom equals to the number of explanatory variables. It will check if any of the regression coefficients are not zero. .05 significance is used.

$H_0$: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

$H_1$: Not all $\beta$'s is 0.

The P-value from the above Omnibus Test table is .000, which is less than the significance level. Hence, rejecting the null hypothesis and accepting the alternate hypothesis of at least one coefficient of predictors is not zero.

**Model Summary**

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|---|---|---|---|
| 1 | 2023.926[a] | .385 | .517 |

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

The Cox & Snell R square of 0.385 and the Nagelkerke R square of 0.517 are similar to the R2 in multiple regression, meaning that these explain the amount of variation in Internet Usage which is predicted by all the 7 independent variables collectively. They're pseudo-statistics because they tend to be r-square, but they're just equivalent to that.

-2 Log Likelihood value is used to assess the model, lower is the value of it better is the model.

**Hosmer and Lemeshow Test**

| Step | Chi-square | df | Sig. |
|---|---|---|---|
| 1 | 5.782 | 8 | .672 |

The Hosmer & Lemeshow test is used to look for the goodness of fit. If the significance level is less than .05, it shows a bad fit. 0.672 indicates support for the model and goodness of fit.

**Variables in the Equation**

| | | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| Step 1[a] | Do you own a mobile phone — yes or no? | 1.676 | .159 | 110.743 | 1 | .000 | 5.347 |
| | Compared to your parents when they were the age you are now, do you think your own standard of living now is better, worse, or about the same as theirs was? | -.003 | .005 | .305 | 1 | .056 | .997 |
| | Do you have access to a working desktop computer, laptop or tablet in your household or at work? | .905 | .083 | 119.451 | 1 | .000 | 2.473 |
| | How old were you at your last birthday? | .070 | .004 | 252.980 | 1 | .000 | 1.073 |
| | Can you read at least some English — yes or no? | .731 | .256 | 8.164 | 1 | .004 | 2.078 |
| | What is the highest level of school you have completed | -.781 | .119 | 42.726 | 1 | .000 | .458 |
| | Urbanity | .324 | .056 | 33.124 | 1 | .000 | 1.383 |
| | Constant | -7.939 | .584 | 184.901 | 1 | .000 | .000 |

a. Variable(s) entered on step 1: Do you own a mobile phone — yes or no?, Compared to your parents when they were the age you are now, do you think your own standard of living now is better, worse, or about the same as theirs was?, Do you have access to a working desktop computer, laptop or tablet in your household or at work?, How old were you at your last birthday?, Can you read at least some English — yes or no?, What is the highest level of school you have completed, Urbanity.

Now, the Individual regression coefficient is evaluated for each predictor in the Variable in the Equation block. Referring to the above coefficients table having p-values to test each regression coefficient.

$H_0$: $\beta_k = 0$

$H_1$: $\beta_k \neq 0$

K= Different Independent Variables.

The p-value in the above "Variables in the Equation table", highlighted in yellow is not significant, as it is greater than the significance level, the rest all the six predictors are significant. We fail to reject the null hypothesis of the highlighted variable and conclude that the coefficient of this variable is not related to Internet Usage and is not an effective predictor. Hence, removing it from the regression equation.

- **Model 2:**

In this model, removing the insignificant predictor (Standard Of Living) from the above model makes no difference in the values of Cox & Snell R square and the Nagelkerke R square. It means the removed predictor variable is not important. Below is a summary of the final model.

**Omnibus Tests of Model Coefficients**

| | | Chi-square | df | Sig. |
|---|---|---|---|---|
| Step 1 | Step | 1120.122 | 6 | .000 |
| | Block | 1120.122 | 6 | .000 |
| | Model | 1120.122 | 6 | .000 |

Global Hypothesis:

$H_0$: $\beta_1 = \beta_2 = 0$

$H_1$: Not all $\beta$'s is 0.

As per the Omnibus Tests of Model Coefficients block shown above, the p-value (.000) is less than the level of significance. Therefore, rejecting the null hypothesis and accepting the alternative hypothesis.

**Model Summary**

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|---|---|---|---|
| 1 | 2024.234[a] | .385 | .517 |

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

The Cox & Snell R square is 0.385 and the Nagelkerke R square is 0.517. All six predictors collectively describe the variation in Internet Usage by 51 percent approximately.

**Hosmer and Lemeshow Test**

| Step | Chi-square | df | Sig. |
|---|---|---|---|
| 1 | 6.240 | 8 | .620 |

Since the Sig. value is greater than .05, it supports the goodness of fit in the final model. Hence, passing this test.

**Variables in the Equation**

| | | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| Step 1[a] | Do you own a mobile phone — yes or no? | 1.678 | .159 | 110.903 | 1 | .000 | 5.353 |
| | Do you have access to a working desktop computer, laptop or tablet in your household or at work? | .905 | .083 | 119.449 | 1 | .000 | 2.472 |
| | How old were you at your last birthday? | .070 | .004 | 253.288 | 1 | .000 | 1.073 |
| | Can you read at least some English — yes or no? | .731 | .256 | 8.161 | 1 | .004 | 2.077 |
| | What is the highest level of school you have completed | -.780 | .119 | 42.602 | 1 | .000 | .459 |
| | Urbanity | .325 | .056 | 33.250 | 1 | .000 | 1.383 |
| | Constant | -7.953 | .583 | 185.916 | 1 | .000 | .000 |

a. Variable(s) entered on step 1: Do you own a mobile phone — yes or no?, Do you have access to a working desktop computer, laptop or tablet in your household or at work?, How old were you at your last birthday?, Can you read at least some English — yes or no?, What is the highest level of school you have completed, Urbanity.

Individual Coefficient Evaluation:

$H_0$: $\beta_k=0$

$H_1$: $\beta_k\neq0$.

As per the above table, the p-values of all the independent variables are significant and are different from zero. All regression coefficients of these predictor variables are related to Internet Usage.

Since the model accuracy is acceptable and all the hypothesis test is passed, making it the final model.

## V. Final Model Summary

Model 2 is the final model and is used for regression analysis in this project.

Please refer to the Omnibus Test, Model Summary, Hosmer And Lemeshow Test, and Variables in the Equation screen capture in Model 2 to have a brief description of the model, and below are the details of each of its elements.

**R Square**

The percentage of the overall difference in the dependent variable Y that the variation in the independent variable X reflects or accounts for. The Cox & Snell R square of 0.385 and the Nagelkerke R square of 0.517 are similar to the $R^2$ in multiple regression

**-2 Log likelihood**

It is similar to the standard error value in Multiple regression. It is 2024.234 for our final model i.e. Model 2 which is explained above in a detailed manner. Less is the value of it, more is the goodness of fit.

**Omnibus Test of Model Coefficients:**

This table is used to test the significance of the model. Hypothesis testing is done and explained in the Model 2 creation which is our final model.

To add to the explanation, chi-square is used to test the significance with degrees of freedom= Number of predictors i.e 6. Using the 0.05 significance level, H0 can be rejected if the calculated Tested statistics is greater than the critical value.

Since, Test Statistics > Critical Value, the null hypothesis (H0) of having regression coefficients as Zero is rejected and an alternate hypothesis (H1) is accepted for this model. Below are the two hypotheses used in this evaluation.

H0: $\beta_1 = \beta_2=0$

H1: Not all β's is 0.

**Classification Table:**

**Classification Table[a]**

| Observed | | | Predicted: Do you currently use the internet — yes or no? Yes | Predicted: Do you currently use the internet — yes or no? No | Percentage Correct |
|---|---|---|---|---|---|
| Step 1 | Do you currently use the internet — yes or no? | Yes | 762 | 225 | 77.2 |
| | | No | 261 | 1054 | 80.2 |
| | Overall Percentage | | | | 78.9 |

a. The cut value is .500

A classification table is used to get the accuracy of the model. PAC, sensitivity, specificity is the extract from this table.

PAC of 78.9 means that, the model has correctly predicted 78.9% of cases. Out of 987 individuals who reported using the internet, the model predicted 762 of them correctly, and out of 1315 individuals who said "No" to using the internet, the model predicted 1054 of them correctly.77.2% and 80.2 are the specificity and sensitivity of the model respectively.

**Variables in the Equation:**

Please see Model 2 (Final Model) for the screen capture of this table. Extending the Explanation, The Wald value is the test statistic similar to the t-statistic in multiple regression. It is used to check the significance of each predictor by comparing it with the critical value. Below are the two hypotheses used in this evaluation.

$H_0$: $\beta_k=0$

$H_1$: $\beta_k\neq0$.

The B values in the table are similar to β coefficients value from multiple regression. The Positive and Negative Sign in it represents the type of relationship of each variable with the dependent variable individually.

Regression Equation is:

**Log(p/1-p) = -7.953 + 1.678(Own a Mobile Phone) + .905(Access to working Desktop at household or at work) + .070(Age)+ .731(Read some English)-.780(Highest Level of Education)+.325(Urbanity)**

Log(p/1-p) is called Log odds. If we increase "Own a Mobile Phone" by a unit, it changes log odds by 1.678. Similarly, if we increase the "Highest level of education" by one unit, it will reduce the log odds by .780 as it is inversely proportional to it. P in log odds is the probability of using the internet and 1-p is the probability of not using the internet.

**Odds Ratio:**

Please see Model 2 (Final Model) for the screen capture of this table. The Exp(B) values for each variable in the table represent the odds ratio of that variable. As the odds ratio increases, the probability of the outcome occurring increases. It can be understood as Odds of using the internet is 5.353 times higher if the individual reports to "Own a Mobile Phone" when all the other factors are equal. If the odds ratio is less than 1, it means that the predictor variable is inversely proportional to the response variable. In this scenario, if "Highest Level of Education" increases by 2 units, Exp(B) value is multiplied twice as 0.459*0.459 = 0.210681 by which will reduce the odds of using the internet.

## VI. Conclusion:

This project focused on applying Logistic Regression to predict whether the individual is using the internet based on multiple factors affecting it. Out of all the reasons that influence the usage of the internet, having a cell phone raises the chances of most users accessing the Internet. The model built is predicting the internet usage on given factors with 51.7 % accuracy.

## VII. References:

[1] X. Chen and R. Ye, "Identification Model of Logistic Regression Analysis on Listed Firms' Frauds in China," 2009 Second International Workshop on Knowledge Discovery and Data Mining, Moscow, 2009, pp. 385-388, doi: 10.1109/WKDD.2009.35.

# Part C: Principal Component Analysis

## I. INTRODUCTION TO PRINCIPAL COMPONENT ANALYSIS

The key aspect of the principal component analysis (PCA) is to minimize the dimensionality of a data set composed of a large number of interrelated variables while preserving as much variance as possible present in the data set. This is done by converting to a new set of variables, the principal components (PCs), which are not correlated and which are arranged in such a way that the first few preserve much of the variance found in all the original variables.

## II. EXPLAINATION

Let's Suppose that x is a vector for k random variables and that the variances of k random variables and the covariances between p variables are of concern. Unless p is small or the structure is very simple, it would always not be very useful to merely look at p variances and all ½*k(k − 1) correlations or covariances. A substitute to it is to search for a few derived variables that hold much of the information provided by these variances and correlations. This is why PCA is performed to get rid of this issue and transform it into a much smaller set of uncorrelated variables called principal components.

The first principal component $PC_1 = a_1 X_1 + a_2 X_2 + ... + a_k X_k$ is the weighted combination of the k observed variables which account for the most variance in the original set of variables, the second principal component is the linear combination which accounts for the most variance in the original variables but being unrelated to the first principal component. Likewise, each further component increases the variance accounted for but is uncorrelated to all the previous components.

## III. DATASET CHOSEN FOR PCA

To illustrate the understanding of PCA (Principal component analysis), a Dataset is chosen from Kaggle (https://www.kaggle.com/davra98/nba-players-20162019) to perform PCA and explain every step of it.

### A. Description

The dataset contains details of NBA players from 2016-2019. It contains 45 columns of which PCA is applied on selective 22 columns.

## IV. STEPS OF PRINCIPAL COMPONENT ANALYSIS WITH EXAMPLE

### A. Select Appropriate Dataset

Multiple requirements need to be considered while selecting a dataset for performing the Principal component analysis on it. These are the size of the sample taken and the intensity of the relationship between the variables.

#### 1) Sample Size

Recommendations for the sample size to achieve stable estimates range from at least 5 observations per variable to 10–20 observations per variable and at least 100 in total. Larger samples are considered better. For smaller sample correlations will not be reliably estimated.

The chosen dataset is having 1408 observations which are fulfilling this requirement.

#### 2) Strength of the relationship among the variables

There should not be a weak relationship between the variables undergoing PCA. To look at the relationship among variables, correlation coefficients are taken into consideration. There should be a significant number of correlation coefficients greater than 0.3 to perform PCA.

For the dataset chosen, maximum correlation coefficients among the variables are greater than 0.3. Below is the correlation matrix among the variables.

#### 3) Bartlett's Test of Sphericity

In this test, a hypothesis test is conducted. It will check if the variables are correlated .05 significance is used.

$H_0$: Correlation between variables is zero.
$H_1$: Correlation exists among variables.

For the studied dataset, P-value is .000 which is less than the level of significance, hence rejecting the null hypothesis and accepting the alternate hypothesis that variables are correlated.

#### 4) Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy

This tests the degree to which the correlation between pairs of variables can be justified by other variables. The maximum value for this is 1. This value should be greater than 0.6 for successful factor analysis.

For the studied dataset, the KMO value is .834, which is greater than 0.6, and according to this PCA can be performed successfully.

Below is the KMO and Bartlett's Test for the dataset used to perform PCA.

**KMO and Bartlett's Test**

| | | |
|---|---|---|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy. | | .834 |
| Bartlett's Test of Sphericity | Approx. Chi-Square | 76640.548 |
| | df | 231 |
| | Sig. | .000 |

### B. Extract the factors.

This step includes assessing the number of factors that best defines the relationship between the variables.

It depends upon the below two factors:

- Rationalizing the number of factors i.e choosing a significant number of factors that can be used.
- Selecting those factors that explain maximum variance.

#### 1) EigenValue

The eigenvalue of a given component calculates the variance in all the variables that are accounted for by that component.

The PCA aims to describe the variation with as few factors as possible, and since each eigenvalue refers to a potentially

different factor, typically only factors with high eigenvalue are kept. The element with the highest eigenvalue has the most variance.

In the below screen capture, for the studied data, the first three principal components are explaining 77% of the variance in the data and are chosen as they have an eigenvalue much greater than 1. The principal components are chosen by passing a fixed number of factors i.e. 3 in the SPSS application. The first principal component has an eigenvalue of 11.025 which explains approximately 50% of the variance in the data. Likewise, if two principal components are considered together, 67 % of the variance is explained by them in the data.

**Total Variance Explained**

| | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | | Rotation Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|---|---|---|
| Component | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 11.025 | 50.115 | 50.115 | 11.025 | 50.115 | 50.115 | 10.029 | 45.585 | 45.585 |
| 2 | 3.910 | 17.773 | 67.887 | 3.910 | 17.773 | 67.887 | 3.505 | 15.931 | 61.516 |
| 3 | 1.987 | 9.030 | 76.917 | 1.987 | 9.030 | 76.917 | 3.388 | 15.401 | 76.917 |
| 4 | 1.054 | 4.791 | 81.708 | | | | | | |
| 5 | .856 | 3.890 | 85.598 | | | | | | |
| 6 | .672 | 3.057 | 88.655 | | | | | | |
| 7 | .613 | 2.787 | 91.443 | | | | | | |
| 8 | .401 | 1.821 | 93.264 | | | | | | |
| 9 | .372 | 1.692 | 94.956 | | | | | | |
| 10 | .352 | 1.598 | 96.554 | | | | | | |
| 11 | .289 | 1.313 | 97.867 | | | | | | |
| 12 | .204 | .929 | 98.796 | | | | | | |
| 13 | .136 | .619 | 99.415 | | | | | | |
| 14 | .077 | .350 | 99.765 | | | | | | |
| 15 | .025 | .114 | 99.879 | | | | | | |
| 16 | .012 | .056 | 99.935 | | | | | | |
| 17 | .008 | .035 | 99.970 | | | | | | |
| 18 | .006 | .026 | 99.997 | | | | | | |
| 19 | .000 | .001 | 99.998 | | | | | | |
| 20 | .000 | .001 | 99.999 | | | | | | |
| 21 | 9.413E-5 | .000 | 100.000 | | | | | | |
| 22 | 6.413E-5 | .000 | 100.000 | | | | | | |

Extraction Method: Principal Component Analysis.

### 2) Kaiser's Criterion:

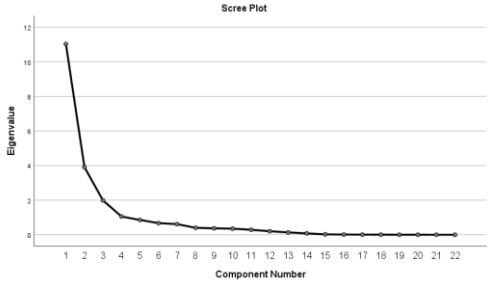
Kaiser's Criterion rule states that components having an eigenvalue greater than 1 are chosen.

As shown above, in the Total Variance Explained table, components having an Eigenvalue greater than one are chosen to have the maximum variance explained by the principal components.

The fourth component is much near to 1 and is not considered in this analysis and also it has an eigenvalue very near to the below component and is not contributing much to the variance explained.

### 3) Catell's scree Plot

Scree plot is a graphical representation of different components which is on the x-axis and their eigenvalues on the y-axis. Factors above the elbow break are considered. Components 1, 2, and 3 are considered as the principal component as per the below plot as these are the components before the elbow break. It is nearly a straight line from 22nd to 4th Component.




## C. Rotate the factors for enhanced interpretation.

Rotation is a method used to make components more explainable without modifying the underlying mathematical properties. These can be achieved by orthogonal (Factors are uncorrelated) and oblique rotation (allowed to correlate).

The varimax rotation (orthogonal rotation) used in this project aims to purify the loading matrix columns such that each component is described by a minimal collection of variables (i.e. each column has a few big loadings and a lot of very small loadings).

The below table shows the relationship of each component with the variables. This table shows the relationship coefficient of 0.5 or greater than that as below it has been suppressed during the analysis in SPSS.

**Rotated Component Matrix[a]**

| | Component | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Field Goals Per Game | .902 | | |
| Field Goal Attempts Per Game | .885 | | |
| Field Goal Percentage | | | .902 |
| 3-Point Field Goals Per Game | | .840 | |
| 3-Point Field Goal Attempts Per Game | | .815 | |
| FG% on 3-Pt FGAs. | | .717 | |
| 2-Point Field Goals Per Game | .917 | | |
| 2-Point Field Goal Attempts Per Game | .932 | | |
| FG% on 2-Pt FGAs | | | .860 |
| Effective Field Goal Percentage | | | .937 |
| Free Throws Per Game | .867 | | |
| Free Throw Attempts Per Game | .897 | | |
| Free Throw Percentage | | .527 | |
| Offensive Rebounds Per Game | .559 | -.532 | |
| Defensive Rebounds Per Game | .802 | | |
| Total Rebounds Per Game | .776 | | |
| Assists Per Game | .678 | | |
| Steals Per Game | .675 | | |
| Blocks Per Game | .537 | | |
| Turnovers Per Game | .879 | | |
| Personal Fouls Per Game | .671 | | |
| Points Per Game | .894 | | |

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 7 iterations.

## D. Interpret the Results

After analyzing the contents of items with high loading from each factor, look at how they match together conceptually, and can be named.

The first component is the different statistics of the basketball player per game can be named "Player's Stats per Game", the second component is related to the 3-point goal statistics of the player can be named "3- Point Goals" and the third component is related to Field Goal percentage of the player can be named as "Field Goal Percentage".

First Component = "Player's Stats per Game"

Second Component = "3-Point Goal"

Third Component = "Field Goal Percentage"

# V. Conclusion

Applying PCA on details of NBA players that were initially spread among 22 variables, reduced to three principal components on different player's statistics.