International Journal of Business Marketing and Management (IJBMM)
Volume 5 Issue 6 June 2020, P.P. 74-87
ISSN: 2456-4559
www.ijbmm.com

Using Data Mining methods to solve classification problems. The connection between the profitability of a financial asset and the profitability of the market portfolio.

Oona Voican[1]
[1]*Economic Informatics Doctoral School, PhD Student*
The Bucharest University of Economic Studies, Romania

ABSTRACT: *Data Mining refers to the analysis of large amounts of data stored in computers. The Big Data era is already present, with current sources indicating that more data have been created over the last two years than they have been generated throughout the entire human history. Big Data involves data sets so large that traditional data analysis methods are no longer usable due to the huge amount of data. Lacking or ignoring the data structure is an extremely important aspect, even more important that size, in data analysis, transformation, innovation and value for business. Data Mining is very effective in many business areas. The key is to find ways or information that can be used in a concrete way to improve business profitability. The Sharpe ratio highlights, by means of an unifactorial model of linear regression, the connection between the profitability of a financial asset and the profitability of the market portfolio, representing an essential step towards the evaluation of primary financial instruments. In other words all assets depend on the market's evolution, but the way they respond to market changes is different.*
KEYWORDS: *data mining, financial asset, market profitability, statistics analysis, sharp regression model.*

I. INTRODUCTION

The usefulness of the Sharp model consists in formulating asset predictions for certain developments in market profitability and portfolio analysis. We will build a Sharp regression model where endogenous variables are represented by the profitability of Oracle and Microsoft. These are our assets of interest, and the independent variable is given by the Standard & Poor's 500 index. The time horizon chosen is 1 year from 25 November 2016 to 24 November 2017, resulting in a sample of 252 daily observations. In order to apply the Sharpe model, however, close prices must be converted into profitability.
The model's equation, obtained after applying the OLS(Ordinary Least Squares) method, is:

$$R_{it} = \propto_i + \beta_i R_{Mt} + \varepsilon_{it} \qquad (1)$$

where i = 1,...,N represents the stocks going into a portfolio and t = 1,...,T describes the time intervals in which the profitability of assets and of the market is tracked.

Thus, the dependent variable R_{it} is given by the historical performances of stock i, while the independent variable R_{Mt} is represented by the historical performances of the market, approximated by the historical performances of a stock index (for instance, the Standard and Poor's 500 index, the Dow-Jones index for the US capital market, the Nikkei index for the stock market in Japan etc.).

The term α_i is a constant, marking the intersection of the estimated line with the Oy axis, without having any kind of economic interpretation. The coefficient β_i - the slope of the model may be interpreted as:

- the profitability response of the i asset to the evolution of the market's portfolio profitability;
- the sensitivity of the i asset to market news;
- the contribution of the i asset to the volatility of the market index.

II. DEFINITION OF TERMS AND CONCEPTS

The desciptive analysis implies the description of the data series, by means of the indicators of the central trend, and also of the variation: **average, median, skewness, kurtosis, standard deviation, quartiles, variation coefficient.** Once transformed into profitability we will import the data into R and we will obtain a data set formed by three variables: Market profitability – Rmarket, Profitability of the Microsoft asset – Rmsft and the Profitability of the Apple asset – Rapple.

The graphical analysis follows the graphical evolution and distribution of variables (**plot, histogram, boxplot, probability density, qqplot** etc.). Since in the current analysis we are particularly interested in profitability distributions, alongside the indicators in bold, we will also verify the Jarque-Bera normality test.

Table 1. Distribution statistics

Indicators	Market profitability	Microsoft asset profitability	Apple asset profitability
First quartile	-0,003846	-0,001681	-0,003664
Median	0,000513	0,000528	0,000976
Average	0,001438	0,000666	0,001622
Third quartile	0,006286	0,002932	0,006939
Standard deviation	0,009561	0,004554	0,010867
Skewness coefficient	1,204102	-0,014312	0,472820
Kurtosis coefficient	9,881057	6,113154	7,645516

Source: Author's assessment

- ✓ Market profitability

The median value is 0.000513, being lower than the average (0.001438), showing a positive skew distribution to the left, with predominantly low values. The average is very close to 0 and positive, indicating that the evolution of market profitability was on average an upward trend. At the same time, on average, market profitability, measured by the Nasdaq Composite stock index, was 0.001438 (0.1438%) over the selected time period.
The standard deviation, which also reflects market risk, implies that, within the chosen time period, market profitability was more than 0.009561 from the average (on average, the returns were higher or lower by 0,009561 than average).
The skewness coefficient is positive, having the value of 1.204102, indicating a positive skew of data series, as the curve, in this case, is slightly elongated to the right, with extremes to the left.
The coefficient of kurtosis = 9.881057 > 3, indicating a leptokurtic distribution, showing that the data have scattered values over a smaller range around the average. Also, the probability of extreme values is higher than in the case of a normal distribution.

- ✓ Profitability of the Microsoft asset

The median value is 0.000528, being lower than the average (0.000666), therefore we have a positive skew distribution to the left, with predominantly small values.

The average is very close to 0 and positive, indicating that the evolution of Microsoft's asset return was on average upward. At the same time, on average, Microsoft's return on assets was 0.000666 (0.0666%).
The standard deviation highlights that, on the chosen time horizon, return on average surpassed 0.000666 from the average (on average, the returns were higher or lower by 0.000666 than the average). The skewness coefficient is negative but close to zero (-0,014312), indicating a weak and negative skew of the data series, the curve, in this case being slightly elongated to the right.

The coefficient of kurtosis = 6,113154> 3, indicating a leptokurtic or high distribution (compared to normal distribution), showing that the data is clustered and close to the average. Also, such a distribution has the thickest "tails", which means that the probability for extreme values is higher than in the case of a normal distribution, which in this case can be explained by the pronounced sensitivity of profitability to various factors influence (situation specific to shares traded on the stock exchange).

- ✓ Profitability of the Apple asset

The median value is 0.000976, being lower than the average (0.001622), showing a positive skew distribution to the left, with predominantly small values.
The average is very close to 0 and positive, indicating that the evolution of Apple's return on assets was on average upward. At the same time, on average, Apple's return on assets was 0.001622 (0.1622%) over the selected time period.

The standard deviation assumes that market profitability was 0.001622 more than the average (on average, returns were higher or lower by 0.001622 than average).
The skewness coefficient is positive, with a value of 0.472820, indicating a positive data series skew, the curve being slightly inclined to the left.
The coefficient of kurtosis = 7.645516> 3, indicating a leptokurtic distribution, indicating that the data has scattered values over a smaller range around the average.

III. METHODOLOGY

The histogram consists of the graphical representation of the frequency distribution of a variable. Analyzing the market yield graph, we notice that it has a slight skew on the left distribution, most values being in the range [-0.01,0.02] with very few observations smaller than – 0.01 or greater than previous. The histogram corresponding to the Microsoft asset's profitability series shows a similar behavior to that of the previous chart. This time, values lower than -0.01 or greater than have the lowest frequency. The last analyzed histogram is for the Apple asset, light skew distribution to the left with with predominantly small values, most values are recorded in [-0.02,0.02].

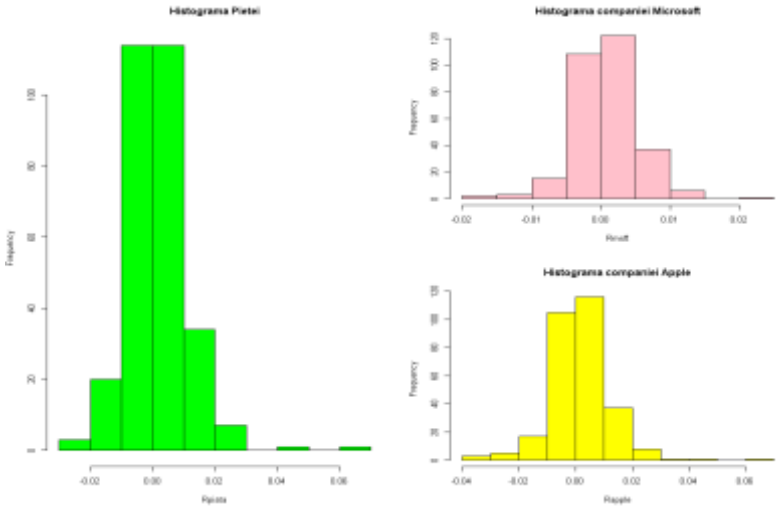

Figure 1. Histogram of the profitability for the market, the Microsoft and Apple assets
Source: Author's assessment

The probability density allows a clearer analysis of time series' distributions. This is a representation of all the values a random variable can take and the probability of occurrence of these values. In the case of a discrete random variable the probability distribution is represented by a function that returns the probability of occurrence for the values of the random variable. In the case of a continuous random variable, the density function is use. Since, for a random continuous variable, the probability of having certain values is 0, the probability that the value of the random variable belongs to a given interval is calculated.

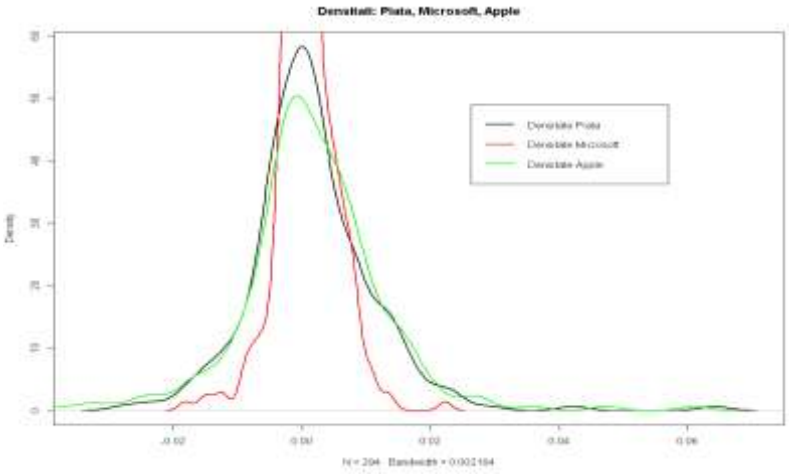

Figure 2. The densities of the three components on a single graph
Source: Author's assessment

The graph for market profitability density confirms the above-mentioned assumptions in the above-described distribution of the market profitability series, according to which the distribution is slightly skew to the left and higher than the normal one. The same assertions are also maintained for the probability density graphs of the profitability of the two assets studied so far.

For a more suggestive illustration of the comparisons between the analyzed and the normal distributions, we will use the ***QQplot*** chart, which allows comparison between two distributions. In the present case, the distribution of each series of the two analyzed corresponds to the normal distribution, in this sense, on the abscissa the normal / theoretical values are quoted, and on the ordinate, the empirical ones, the values of the two series being previously standardized. We use such a QQnorm graph. If the yields come from a normal distribution, then there

will be a linear relation between the series of empirical quantifications and the series of theoretical quantifications.

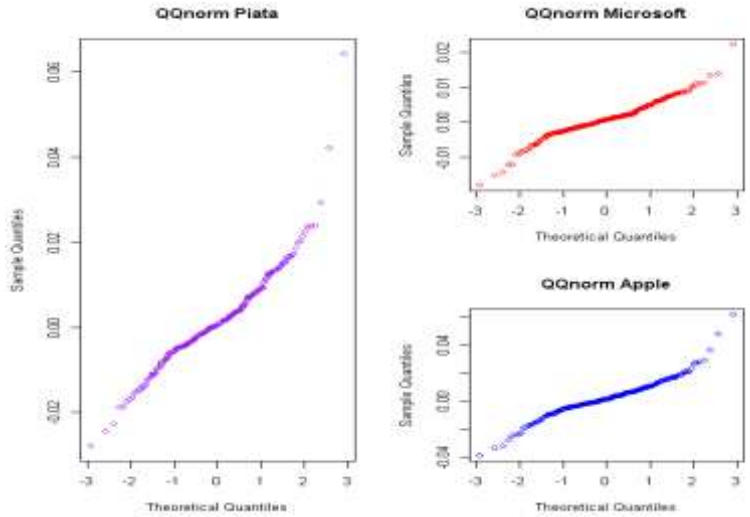

Figure 3. Qqnorm Plot for the analyzed series
Source: Author's assessment

Comparing the distribution of the market profitability series with the normal distribution, we observe that there are similarities, but more values that deviate from the Gauss distribution are present.
It is rather a heavy-tailed distribution, as deviations are very pronounced. Comparing the distribution of the assets of Oracle and Apple to the normal one, we note that the above observations are maintained.

3.1 The Jarque-Bera test

The normality of the series'distributions can also be checked with the Jarque-Bera test. The Jarque-Bera test considers both the skewness and kurtosis coefficients and verifies to what extent the empirical distribution can be approximated with a normal distribution. The null hypothesis of this test assumes that the data sample comes from a normal distribution, and the test statistic is determined as follows:
where, N is the number of observations in the sample, S – skewness coefficient, K – kurtosis coefficient.

The test assumptions are:

- H0: The sample follows a normal distribution
- H1: The sample does not follow a normal distribution.

Table 2. The Jarque-Bera test

Jarque-Bera test for normality
Rpiață JB = 651.07, p-value < 2.2e-16
Rmsft JB = 118.73, p-value < 2.2e-16
Rapple JB =275.32, p-value < 2.2e-16

Source: Author's assessment

We note that in both cases, the p-value (the probability of being wrong in rejecting the null hypothesis) is very close to 0, which means that we will accept the alternative hypothesis that the distributions of the two series are

not normal. Before testing the stationarity of the series and estimating the model, we will check whether there is a connection between the two series, using in this sense the **correlation coefficient.**
In the present case, the correlation coefficient between the Rmarket and Rmsft is 0.5930474, between the Rmarket and Rapple is 0.4640943, and between Rmsft and Rapple is 0.4957914, indicating a direct, strong link between the two variables.

The type of the relationship is given by the sign of the calculated coefficient, highlighting the fact that the related variables vary in the same sense (when one grows, the other grows too, respectively when one decreases, the other decreases).The numeric value provides information about the intensity of the link, the result obtained indicating a strong link. The relationship between the two variables can be graphically represented using a scatter plot, as follows:

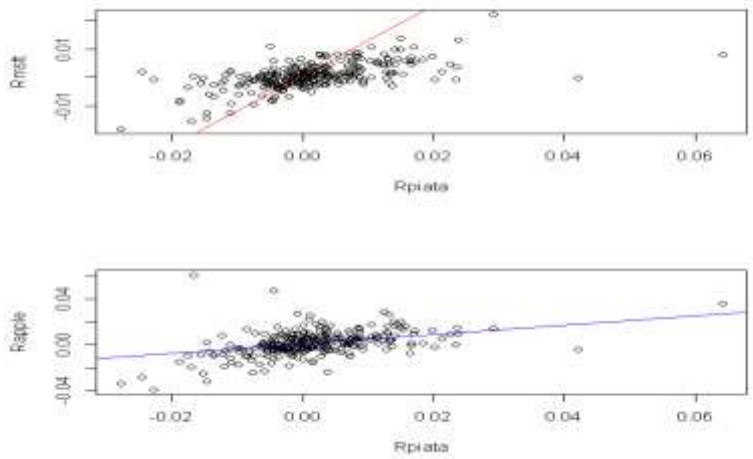

Figure 4. Scatter-Plot
Source: Author's assessment

The visual analysis of the organization and shape of the points cloud obtained can provide important clues regarding the relationship between the variables studied allowing the possibility to establish both the existence of the link between two variables and their meaning and intensity, if present. The horizontal axis represents the independent variable (explanatory) – the profitability of the market and the vertical axis represents the profitability of the Microsoft/Apple asset. In the example considered, we note that the values are distributed around an upward right so, as the profitability of the market increases, the profitability of the asset increases. The line drawn on the graph is the regression line that gives the tendency of the relationship and best approximates the variation of the pairs of values. The trajectory of this line is fixed on the basis of a mathematical model called the "smallest square method", which ensures minimizing the distances between the actual points. The slope of the regression right is of significant interest, as it provides information about the type and intensity of the link by its numerical sign and value.

3.2 Series' Stationarity

Before estimating the regression model, an important step is to check the series' stationarity. In economy, most of the economic series on important indicators (prices, exchange rate, consumption, export, etc.) show a tendency. Such series are considered non-standard series. The fact that the chronological series terms generally have growth or decrease trends over time, makes the average string of YT's values differ depending on the time t from which we think the series begins. Moreover, even the dispersion and covariance are dependent on the variable time t. A stationary series is a series whose value oscillate more or less randomly, around a reference level – the average, thus being in a steady state [7].

Modeling of non-standard series can lead to "false regressions" for which R2 is very high (close to 1), and DW statistic is very low (tends to 0, errors being related), since the series used, being non-stationary, behaves as a random run process (have unit root). In addition, the regression analysis based on chronological series assumes that the time series are stationary. Classic T and F tests are based on this assumption. A simple test of the stationarity of the series that gives us first information in this respect, is based on the function of autocorrelation (ACF). The graph of the autocorrelation function in relation to the K offset is called the correlogram.

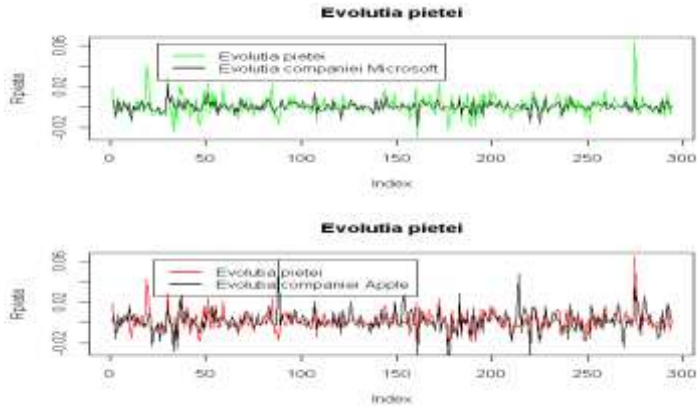

Figure 5. The graphically representation of the two time series using the plot.ts() function.
Source: Author's assessment

We note that the series varies around a reference level close to the value 0, which gives us an indication that the series would be stationary. In the correlogram Fig. 6 corresponding to the series of market profitability, on the abscissa are represented lags and on the ordinate, autocorrelation coefficients.

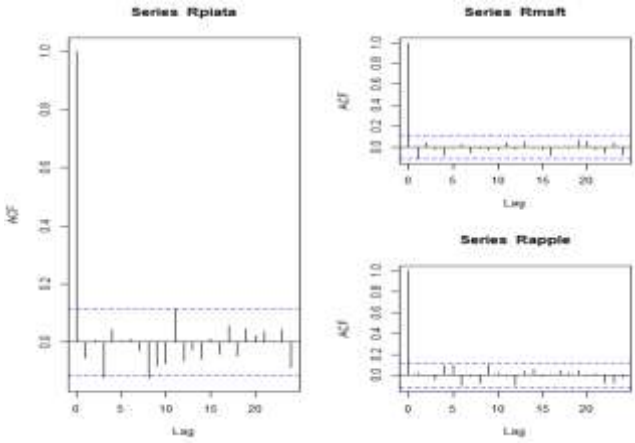

Figure 6. Correlogram – Rpiață, Rmsft, Rapple
Source: Author's assessment

To make sure the series are stationary we will use the ADF test, the assumption of which are:

- H0: The series has unit root and is non-stationary ($\rho = 1$);
- H1: Series is stationary ($\rho < 1$).

```
        Augmented Dickey-Fuller Test

data:  Rpiata
Dickey-Fuller = -6.6791, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary
```

```
        Augmented Dickey-Fuller Test

data:  Rmsft
Dickey-Fuller = -6.9746, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary


        Augmented Dickey-Fuller Test

data:  Rapple
Dickey-Fuller = -6.151, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary
```

Figure 7. Augmented Dickey-Fuller Test
Source: Author's assessment

We note that in all three cases p-value = 0.01 < 0.05 (the chosen materiality threshold), which means that we reject the null hypothesis and accept the alternative that the series are stationary. As the series are stationary, there is no need to differentiate them, thus we can continue with the estimation of the regression model.

IV. ESTIMATION OF THE COEFFICIENTS OF THE SHARP REGRESSION MODEL

Using the lm() function we obtain the next output:

```
Call:
lm(formula = Rmsft ~ Rpiata)

Residuals:
      Min        1Q    Median        3Q       Max
-0.0122363 -0.0022946  0.0000108  0.0020180  0.0137354

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0002598  0.0002167   1.199    0.232
Rpiata      0.2825069  0.0224458  12.586   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.003674 on 292 degrees of freedom
Multiple R-squared:  0.3517,    Adjusted R-squared:  0.3495
F-statistic: 158.4 on 1 and 292 DF,  p-value: < 2.2e-16


  Call:
  lm(formula = Rapple ~ Rpiata)

  Residuals:
        Min         1Q     Median         3Q        Max
  -0.027690 -0.004240 -0.000117  0.004275  0.068848

  Coefficients:
               Estimate Std. Error t value Pr(>|t|)
  (Intercept) 0.0008632  0.0005687   1.518     0.13
  Rpiata      0.5274869  0.0589173   8.953   <2e-16 ***
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Residual standard error: 0.009643 on 292 degrees of freedom
  Multiple R-squared:  0.2154,    Adjusted R-squared:  0.2127
  F-statistic: 80.16 on 1 and 292 DF,  p-value: < 2.2e-16
```

Figure 8. The regression model between Rmsft~Rpiață și Rapple~Rpiață
Source: Author's assessment

The estimated regression model is:

$$Rmsft = 0.0002598 + 0.2825069 * Rpiață + \varepsilon_{it}$$
$$Rapple = 0.0008632 + 0.5274869 * Rpiață + \varepsilon_{it}$$

Further we will check the significance of the parameters using the T-Test:

- H0: $\beta 0 = 0$ (free term is not statistically significant)
- H1: $\beta 0 \neq 0$ (free term is statistically significant)

Whereas the probability associated with the free term to err in rejecting the null hypothesis (Prob. = 0.232 and Prob. = 0.13) is greater the chosen materiality threshold, 5% $\Rightarrow$ accept H0 and reject the alternative, thus the free term is not statistically significant.

Since the associated probability for the slope coefficient, relating to the likelihood of failure in the rejection of the null hypothesis is much lower than the chosen materiality threshold, of 5% $\Rightarrow$ we reject H0 $\Rightarrow$ we accept H1 $\Rightarrow$ the parameter is statistically significant. Therefore, we can interpret it, thus the regression coefficient complements the results offered by the correlation coefficient and shows that the link between the two variables is a direct one ($\beta_1 > 0$). It can also be argued that the increase by one unit (by 1%) of market profitability, the profitability of a stock in Microsoft increases by 0.283 units (by 0.283%), and the increase by one unit (by 1%) of market profitability, the profitability of a stock in Apple increases by 0.527 units (by 0.527%). At the same time, a β_1 subunitary to 1 shows the analyzed asset is less risky than the market (a lower volatility in relation to the market).

Also, for the coefficient of the regression model we can determine the confidence interval, using the cofint() function, illustrated below:

```
> confint(model1, level = 0.95)
                    2.5 %       97.5 %
(Intercept) -0.0001666524 0.0006862122
Rpiata       0.2383309161 0.3266829299
> confint(model2, level = 0.95)
                    2.5 %      97.5 %
(Intercept) -0.0002560971 0.001982564
Rpiata       0.4115305461 0.643443327
```

Figure 9. The coefficients of the regression model
Source: Author's assessment

We note that the free term is not significantly different from 0, as the confidence interval for it comprises the value 0. Given a confidence coefficient of 97.5%, the range [-0.0001666524; 0.0006862122] and the range [-0.0002560971; 0.001982564] will include the actual value of β_0.
For parameter β1 we note that the built-in range does not contain the value 0, indicating that the parameter is significantly different from 0, and we can also note that it records positive values. Given a 97.5% confidence coefficient, in the long term, in 97.5 of 100 cases, the range [-0.2383309161; 0.3266829299] and the [-0.4115305461; 0.643443327] range will include the actual value of β_1.
Returning to the regression model, we note that we have a coefficient of determination (multiple R-squared = 0.3517 and multiple R-squared = 0.2154), showing that approximately 35.17% of the change in the profitability of Microsoft asset and about 21.54% of the change in the profitability of the Apple asset is explained by the profitability of the market, the rest of the variation being explained by other factors that are not included in the model. The adjusted determination coefficient namely R-squared (Model1 = 0.3495 and Model2 = 0.2127) also considers the number of observations and the exogenous variables. Adjusted R2 provides a statistically significant result because it penalizes the introduction of independent variables that have a low relevance in explaining dependent variables.

4.1 Verification of the assumption of the simple linear regression model

- ✓ A first hypothesis relates to the normality of the distribution of random errors and their average.

A first clue about the distribution of the residue series (obtained using the resid() function) is given by the histogram of the random error series illustrated in Fig. 10. It can easily be noticed that the distribution is not a normal one, but has strong skew to the left.

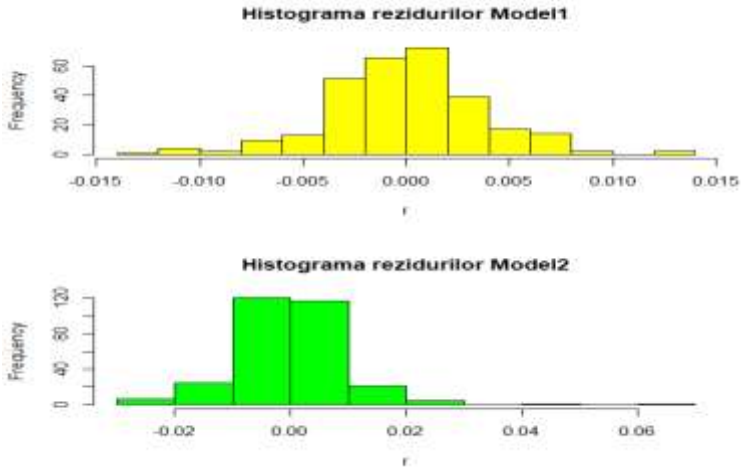

Figure 10. Histogram of the random error series
Source: Author's assessment

However, in order to know with certainty that the residue series is not normally distributed, we will use the Jarque-Bera test, with assumptions:

- H0: Random errors have normal distribution
- H1: Random errors do not have normal distribution.

```
> jb.norm.test(r1,nrep=2000)

        Jarque-Bera test for normality

data:  r1
JB = 23.736, p-value = 0.001

> jb.norm.test(r2,nrep=2000)

        Jarque-Bera test for normality

data:  r2
JB = 1483.8, p-value < 2.2e-16
```

Figure 11. Jarque-Bera test
Source: Author's assessment

It is noted that the P-value is extremely small, which means that we reject the null hypothesis and accept the alternative that the series is not stationary. Therefore, the hypothesis of normality of random errors is not respected.

✓ Homoscedasticity of random errors

In order to establish that random errors are homoscedastic (the property of errors to have a constant variance) or not, we will apply the Breusch-Pagan test with assumptions:

- H0: There is homoscedasticity
- H1: There is heteroscedasticity

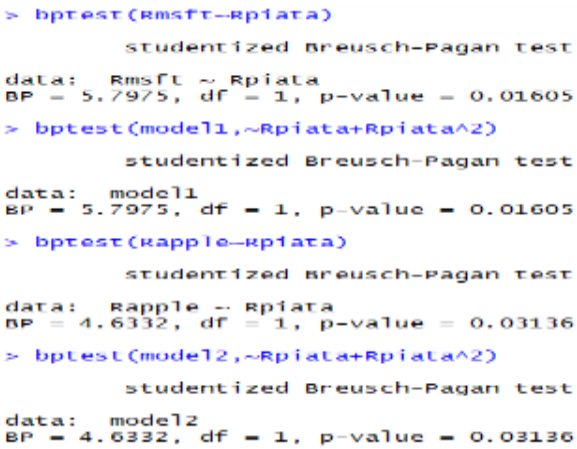

Figure 12. Homoscedasticity
Source: Author's assessment

The value of P-value, in this case is 0.03136, so we can reject the null hypothesis, with a sufficiently high certainty, because P-value must be less than 0.2, and to accept it, P-value must exceed the value of 0.7.

- ✓ Non-autocorrelation of random errors

The Durbin-Watson test checks whether there is first-order autocorrelation in the residue series with assumptions:

- H0: No autocorrelation of the order I
- H1: There is autocorrelation of the order I

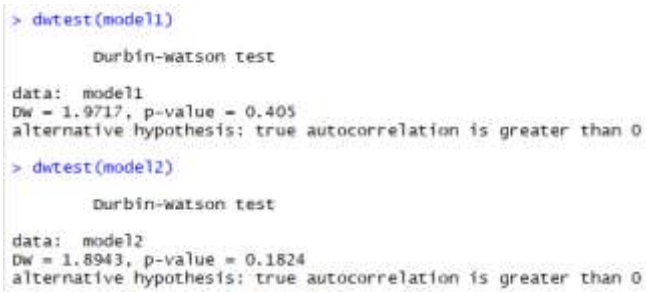

Figure 13. Autocorrelation of the order 1
Source: Author's assessment

As the value of P-value = 0405 and 0.1824 < 0.7, the probability of being wrong in rejecting the null hypothesis is very low, therefore we reject H0, according to which there is no autocorrelation of order I.

To detect random upper-order errors, we use the Breusch-Godfrey test, with assumptions:

- H0: There is no autocorrelation of random errors
- H1: There is autocorrelation of random errors

As P-value = 0.9925 and 0535, as shown in the fig. 14, we reject H0, we accept H1, there is autocorrelation of random errors.

```
> ?"bgtest"
> bgtest(model1, order=3, type = "Chisq")

        Breusch-Godfrey test for serial correlation of order up to 3

data:  model1
LM test = 0.094399, df = 3, p-value = 0.9925

> bgtest(model2, order=3, type = "Chisq")

        Breusch-Godfrey test for serial correlation of order up to 3

data:  model2
LM test = 2.1843, df = 3, p-value = 0.535
```

Figure 14. Autocorrelation of the order 3
Source: Author's assessment

The linearity of the model can be tested using the Ramsey test, with assumptions:

- H0: The model is linear
- H1: The model is not linear

```
> ?"resettest"
> resettest(model1, power = 2:3)

        RESET test

data:  model1
RESET = 8.2294, df1 = 2, df2 = 290, p-value = 0.000334

> resettest(model2, power = 2:3)

        RESET test

data:  model2
RESET = 7.8693, df1 = 2, df2 = 290, p-value = 0.0004698
```

Figure 15. Linearity tested using the Ramsay test
Source: Author's assessment

As the value of P-value = 0.000334 and 0.0004698 < 0.7, the probability of being wrong in rejecting the null hypothesis is very low, so we accept H1, according to which the model is nonlinear.

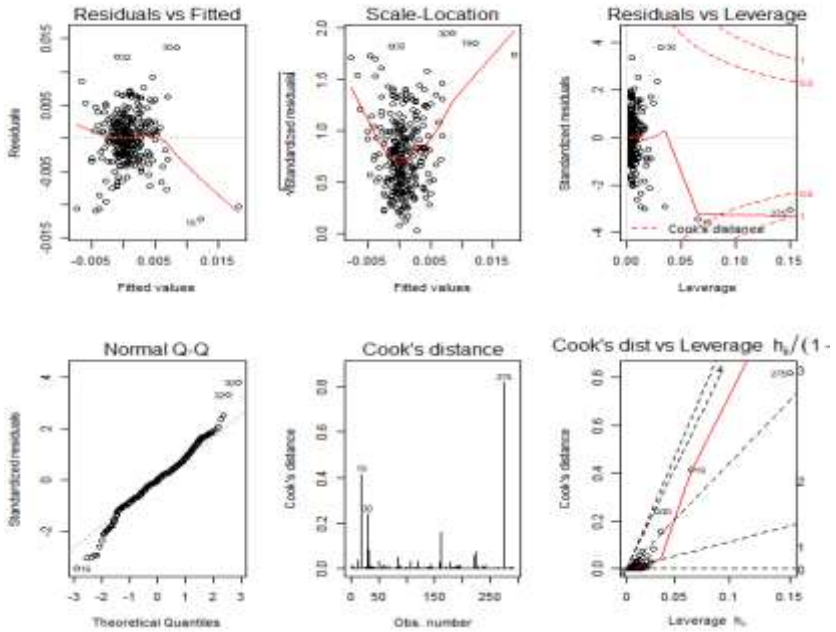

Figure 16. Diagnostic charts Model 1
Source: Author's assessment

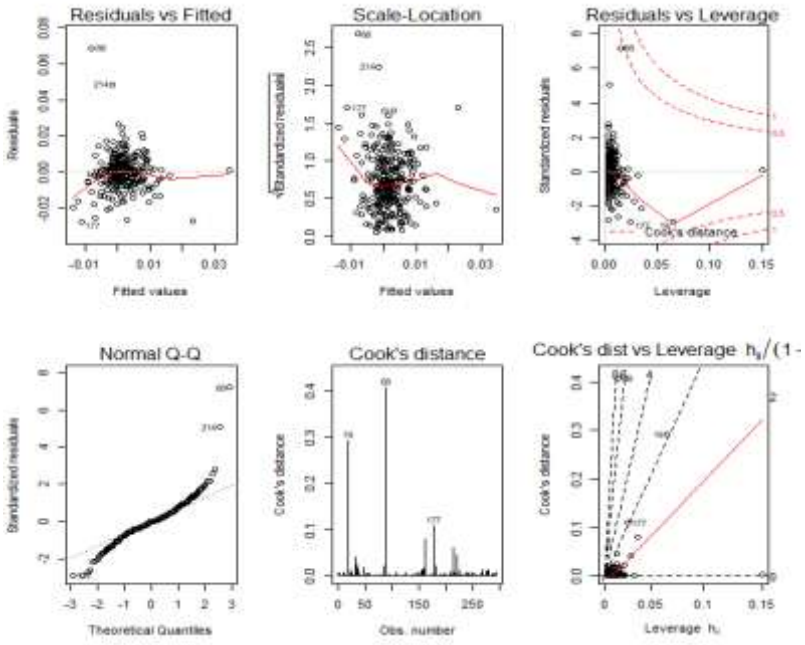

Figure 17. Diagnostic charts Model 2
Source: Author's assessment

Residues' graph vs. forecasted values check the linearity of the model, so the red line should be roughly horizontal, without curvature, which is respected if we analyze the first graph in Fig. 16. The Normal Q-Q diagram verifies the normality of the residues' distribution (normal errors). Points should fall on a diagonal line, but as is observed, there are many values that do not respect this settlement.

The Scale–Location diagram verifies the homoscedasticity hypothesis of errors. In the ideal case, the line should be horizontal, without curvature, and points scattered at a relatively equal distance from the red line. We can observe that this hypothesis is respected to some extent.

The Residuals vs. Leverage diagram helps us discover those observations with major influence on the estimated model. Although there are extreme values, they may not have an impact on the regression line, which means the results would not be very different should we eliminate them from the analysis. On the other hand, there may be values that can greatly influence the regression coefficients and their elimination is recommended. This time, we are interested in the values in the upper right corner or the lower left corner of the graph, areas where observations can be found with high influence. When an observation exceeds Cook's distance (i.e. red dotted lines) means that they have a major influence and is preferable to be removed from the analysis. In the present case, we note that the 116 observation is in the lower left corner, but does not exceed the range [-0.5; 0.5], corresponding to the Cook distance, which means that the value does not have a major influence on the estimated model and can be preserved in the analysis.

V. CONCLUSION

The objective of this paper was to perform two analyzes based on the Sharpe model to determine how two assets: Microsoft's actions and Apple's actions react to market changes.

Before estimating a regression model based on the data series, an important step is to check the series' stationarity. To verify the stationarity of the data series we analyzed the autocorrelation coefficients that are

statistically insignificant starting with lag 1 which demonstrates that the analyzed series are stationary. To make sure the series are stationary we also use the ADF test, which shows that the three series are stationary.

Two regression models were further estimated: in the first model, it was analyzed Microsoft's asset response to market changes, while in the second model it was analyzed Apple's asset response to market changes. The tests performed show that the two variables are positively correlated to a market profitability trend, the increase by one unit (by 1%) of market profitability, the profitability of a stock in Microsoft increases by 0.283 units (by 0.283%), and the increase by one unit (by 1%) of market profitability, the profitability of a stock in Apple increases by 0.527 units (by 0.527%).

Examining the regression model, we note that we have a coefficient of determination (multiple R-squared = 0.3517 and multiple R-squared = 0.2154), showing that approximately 35.17% of the change in the profitability of Microsoft asset and about 21.54% of the change in the profitability of the Apple asset is explained by the profitability of the market, the rest of the variation being explained by other factors that are not included in the model.

REFERENCES

[1] E.L. Allwein, R.E. Schapire, and Y. Singer, Reducing multiclass to binary: a unifying approach for margin classifier. *Journal of Machine Learning Research, 1*, 2001, 113-141.

[2] M. Balkin, and P. Niyogi, Semi-supervised learning on Riemannian manifolds. *Machine Learning, Special Issue on Clustering, 56*, 2004, 209-239.

[3] M. Belkin, P. Niyogi, and V. Sindhwani, Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research, 7*, 2006, 2399-2434.

[4] L. Cai, and Y. Zhu, The challenges of data quality and data quality assessment in the big data era. Interactive:http://datascience.codata.org/articles/10.5334/djs-2015-002/print/, 2015.

[5] L. Canzian, and M. van der Schaar, Real-Time Stream Mining: Online Knowledge Extraction Using Classifier Networks, *IEEE Network 29, no. 5*, pp. 10– 16.

[6] K.N. Cukier, and V. Mayer-Schoenberger, *The Rise of Big Data: How It's Changing the Way We Think About the World. In The Fourth Industrial Revolution* (A Davos Reader, ed. G. Rose, Chapter 3, 2013 from Foreign Affairs).

[7] T.H. Davenport, and J. Dyche (2013). Big Data in big companies. International Institute for Analytics, 2013.

[8] T.H. Davenport, *Big Data at Work* (Boston, MA: Harvard Business Review Press, 2014)

[9] M. Evans, Healthcare Data Mining, *Modern Healthcare 45, no. 39*, 2015, p. 24.

[10] R. Flamary, D. Tuia, B. Labbe, G. Camps-Valls, and A. Rakotomamonjy, Large margin filtering. *IEEE Transactions on Signal Processing, 60(2)*, 2012, 648-659.

[11] M. Fuller, *Big data: New science, new challenge, new dialogical opportunities* (Zygon, 50(3), 2015, 569-582).

[12] L. Gomez – Chova, G. Camps-Valls, L. Bruzzone, and J. Calpe-Maravilla, J. Mean map kernel methods for semisupervised cloud classification. *IEEE Transactions on Geoscience and Remote Sensing, 48(1),* 2010, 207-220.

[13] I.A.T Hashem, I.Yaqoob, N.B Anuar, S. Mokhtar, A. Gani, and S.U. Khan, The rise of "big data" on cloud computing: Review and open research issues. *Information Systems, 47*, 2015, 98-115.

[14] T. Hastie, R. Tibishirani, and J. Friedman, *The Elements of Statistical Learning (*Springer-Verlag, New York, NY., 2001)

[15] J. Heskett, How will the "Age of Big Data" affect management? HBS Working Knowledge, 2012).