2016 WMI Student Seminars

"The Rising Importance of Mathematical Statistics in Economic Fields: The Stock Market"

Yongjun Lee (ylee2021@chadwickschool.org)
July 2016

1. Introduction

In this paper, we will be exploring the behavior of the stock market in terms of mathematics. We can analyze this by applying mathematical techniques to an organized pool of data, a procedure known as "mathematical statistics". Some of these techniques include regression analysis (using trends to establish a broad picture of the data), distribution analysis (determining how widely spread the data is), and statistical representations (finding a single number that represents the whole data set). Today's presentation will mostly involve regression analysis as the approved method of analyzing large quantities of data.

2. Exponential Growth/Trends

When creating a general pattern or algorithm for a large set of data, a phenomenon that often arises in such cases is exponential growth. This occurs when the y-variable increases by a constant factor larger than 1 when the x-variable increases by 1. This factor is known as the base or growth factor.

One might think that this sort of graph is bizarre, but this actually occurs in an extremely high percentage of real-life situations, due to the exponential behavior of life forms. When organisms such as bacteria and humans grow, each bacterium or human has a certain number of offspring, which then also have the same number of offspring, and so on. This creates an exponential trend wherein the number of offspring for each parent becomes the growth factor. Other non-living systems, such as bank accounts and nuclear chain reactions, also show this kind of trend.

Another concept that needs explanation prior to analysis is the R^2 value. This number represents the strength of a particular trend; the higher it is, the stronger the trend will be. If an R^2 value is below 0.5, then its correlation is considered weak. An R^2 value between 0.5 and 0.8 indicates a normal correlation, whereas anything above 0.8 is the sign of a strong correlation.

Furthermore, there are some technical nuances that make interpretation a little difficult. This is because Excel does not accept dates, and instead transforms them into serial numbers. In Excel's system, January 1, 1900 is designated as 1, January 2, 1900 is 2, January 3, 1900 is 3, and so on. Therefore, an increase of 1 in the serial number is equivalent to a day passed.

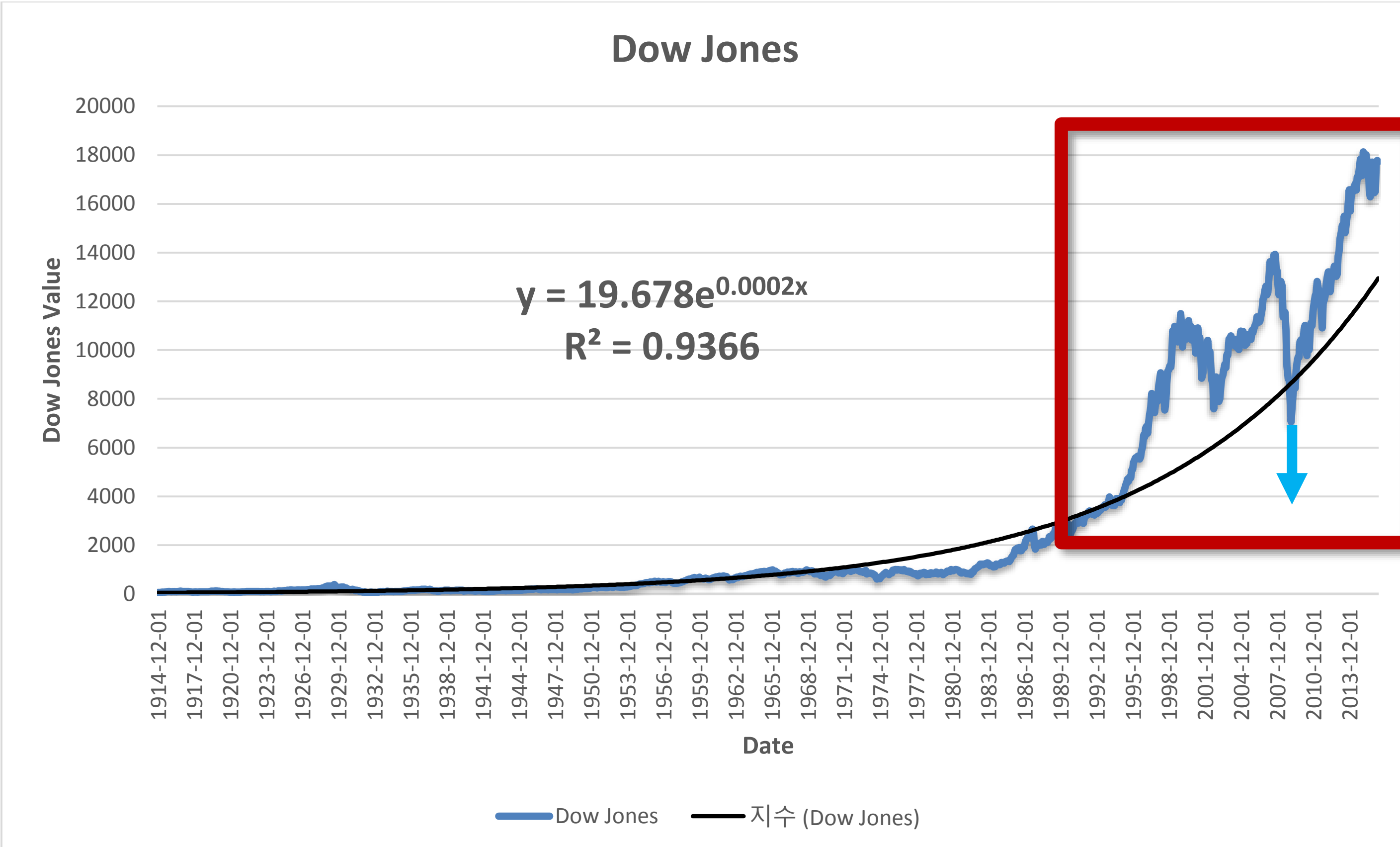
3. Analysis I: Total

In this analysis, we will use two different data sets, both from reputable stock market indexes. A stock market index is a value that attempts to measure the stock market as a whole by looking at several representative companies. Here, we will be looking at the Dow Jones and the NASDAQ, which some of you might have heard about in the news or other media.

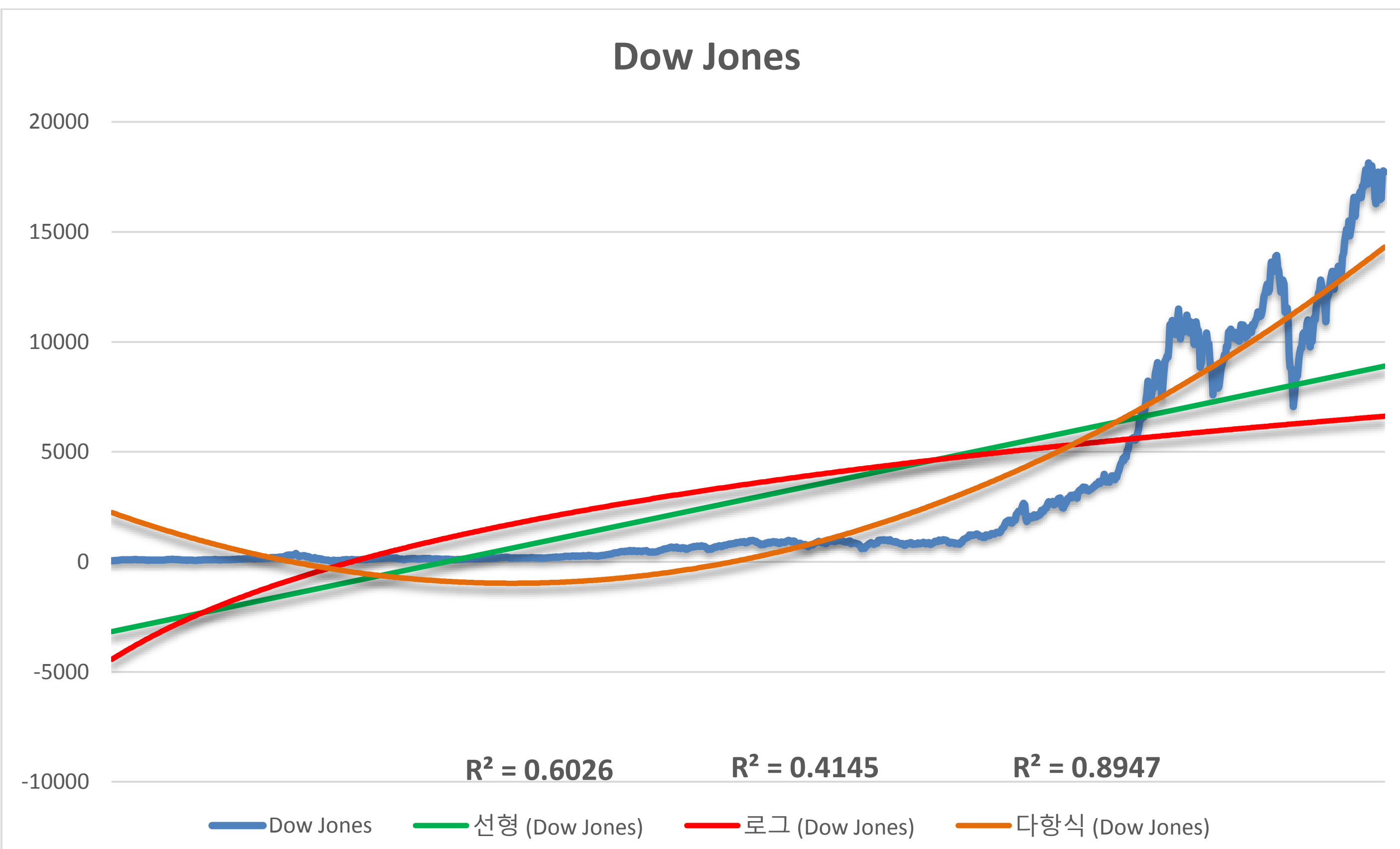
One thing to note specifically for this analysis is that we focus on the whole graph for both stock indexes, instead of a certain time period. This allows us to look at the graph to its fullest by determining trends for the entire data set. It also prepares us for our next analysis, which will break the data into chunks; this way, we can pinpoint and later analyze particular moments in time where the correlation is very strong or becomes weak.

i. Dow Jones: 1905-Present

In the Dow, we have a century's worth of data; a full 1200 data points, one for each month. As shown in the image directly below, the trend shows an exponential graph with the exponential factor being $e^{0.0002} \approx 1.0002$ in this case. In other words, the Dow increased by a factor of approxiately 1.0002 each day. The strength of this correlation can also be seen from the second picture; the exponential trend's R^2 value of 93.66 is higher than any other, including the linear and polynomial trendlines. Furthermore, we can also see that the projected trend ends far below the current index value; this can be interpreted as an acceleration of the speed with which the index increases. Another important portion of the graph to note here is the sharp drop from 2007 to 2008 during the global financial crisis, shown here by the blue arrow. Furthermore, as shown in the red frame, this seems to have highly affected the trend, driving it far down below most of the other values from the last two centuries.



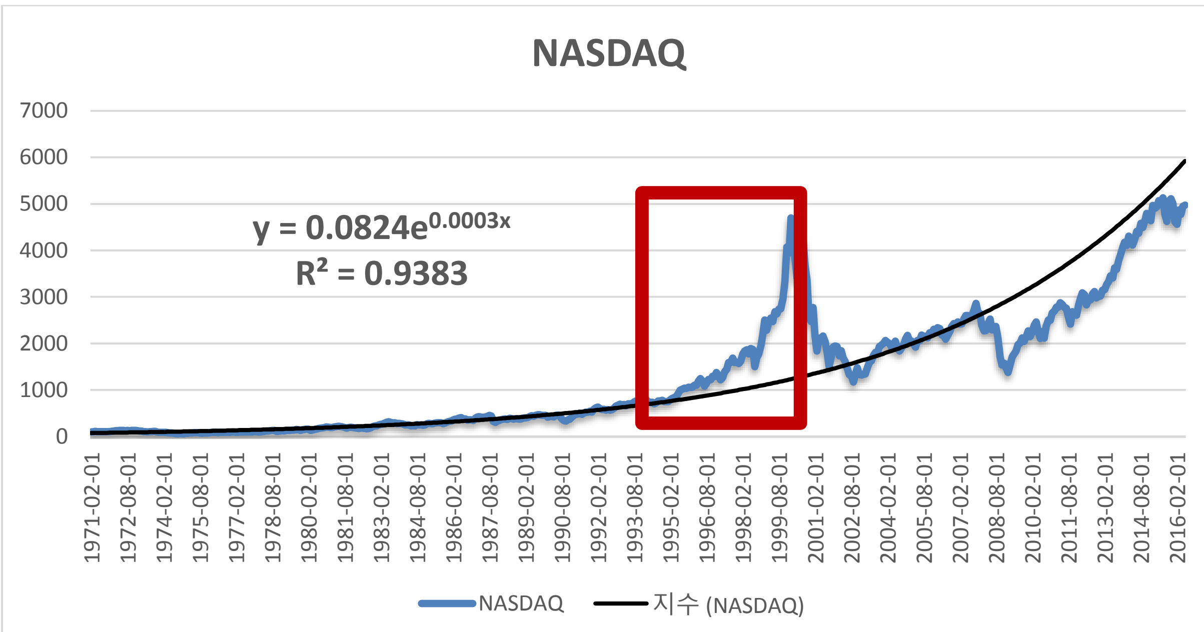
Dow Jones
Dow Jones Value
Date
y = 19.678e^0.0002x
R² = 0.9366
20000
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
1914-12-01
1917-12-01
1920-12-01
1923-12-01
1926-12-01
1929-12-01
1932-12-01
1935-12-01
1938-12-01
1941-12-01
1944-12-01
1947-12-01
1950-12-01
1953-12-01
1956-12-01
1959-12-01
1962-12-01
1965-12-01
1968-12-01
1971-12-01
1974-12-01
1977-12-01
1980-12-01
1983-12-01
1986-12-01
1989-12-01
1992-12-01
1995-12-01
1998-12-01
2001-12-01
2004-12-01
2007-12-01
2010-12-01
2013-12-01
Dow Jones
지수 (Dow Jones)

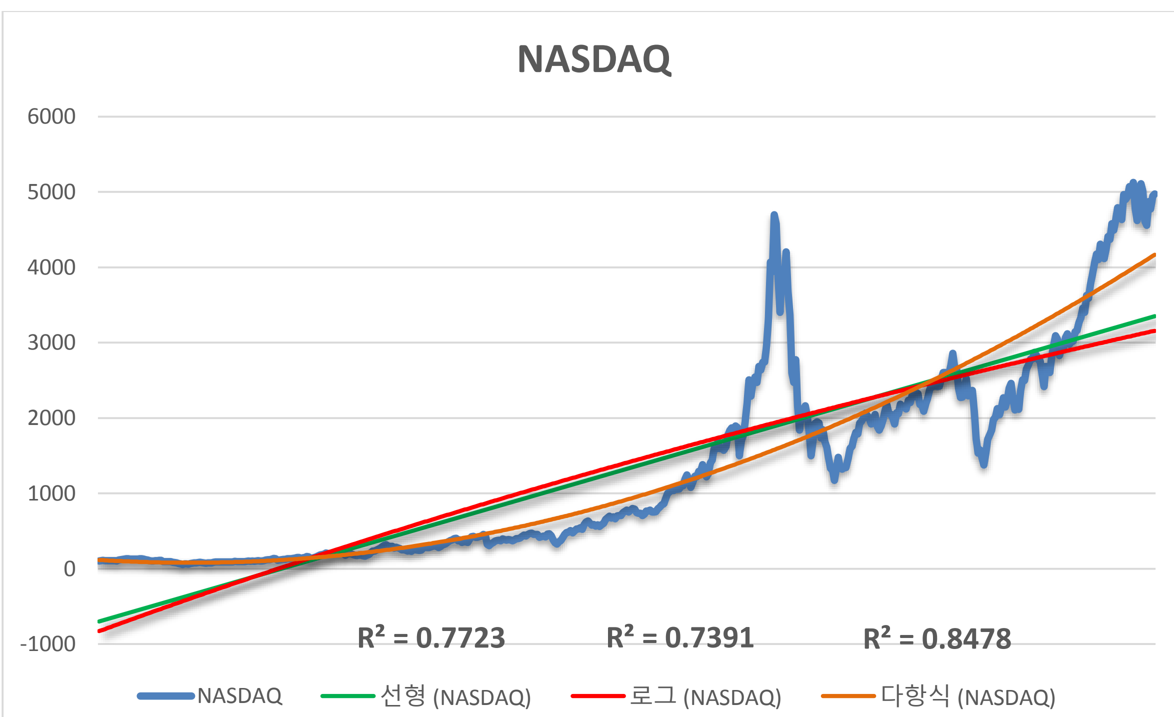


Dow Jones
20000
15000
10000
5000
0
-5000
-10000
R² = 0.6026
R² = 0.4145
R² = 0.8947
Dow Jones
선형 (Dow Jones)
로그 (Dow Jones)
다항식 (Dow Jones)

ii. NASDAQ: 1971-Present

Meanwhile, in the NASDAQ, a more recently published but equally trustworthy stock index, we analyze approximately 540 data points. In this trend, the exponential factor is higher ($e^{0.0003} > e^{0.0002}$), because the region from 1915 to 1970, when there is little increase in the data, has been omitted. However, the graph still shows an exponential trend, even stronger than the Dow, with a slightly higher R^2 value of 0.9383. Here in the graph, we can note the sharp increase from 1994 to 2000 as the most prominent feature, as framed here in red. This increase also seems to be exponential, giving this portion more promise in terms of trends.





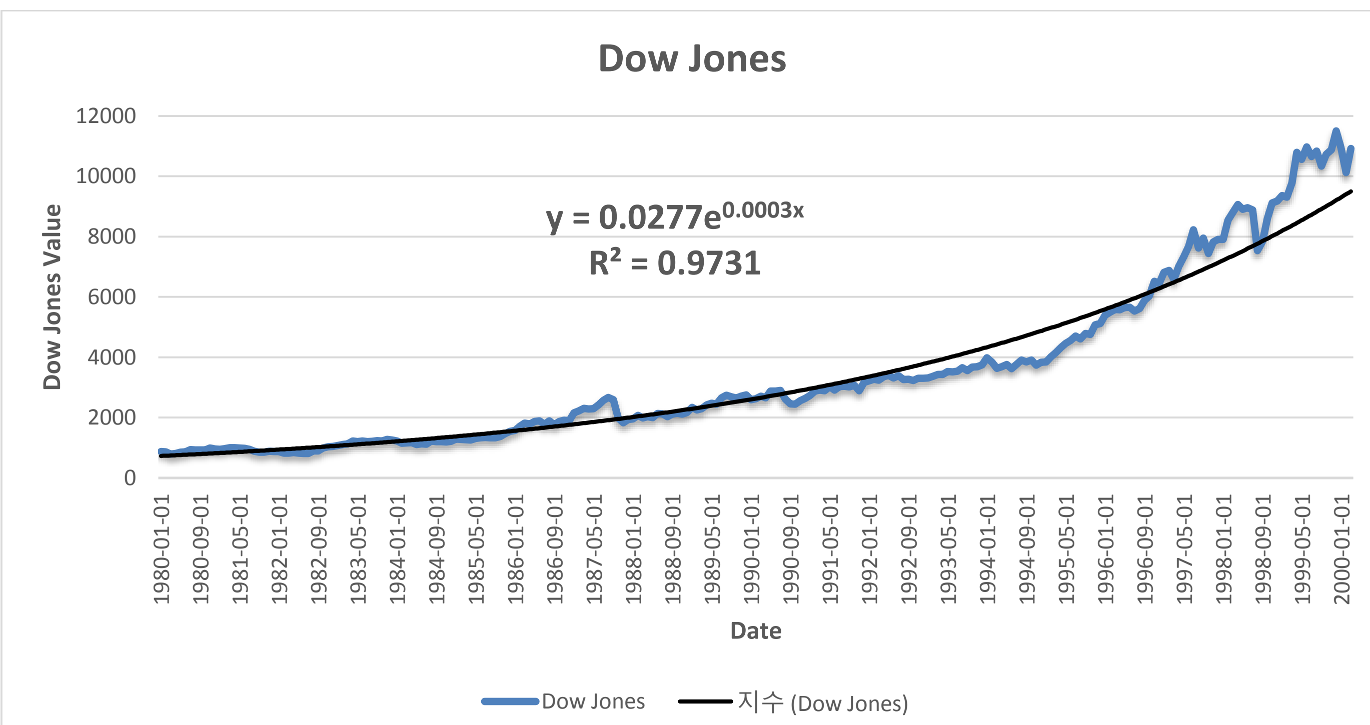
4. Analysis II: The “Dot-Com Bubble”

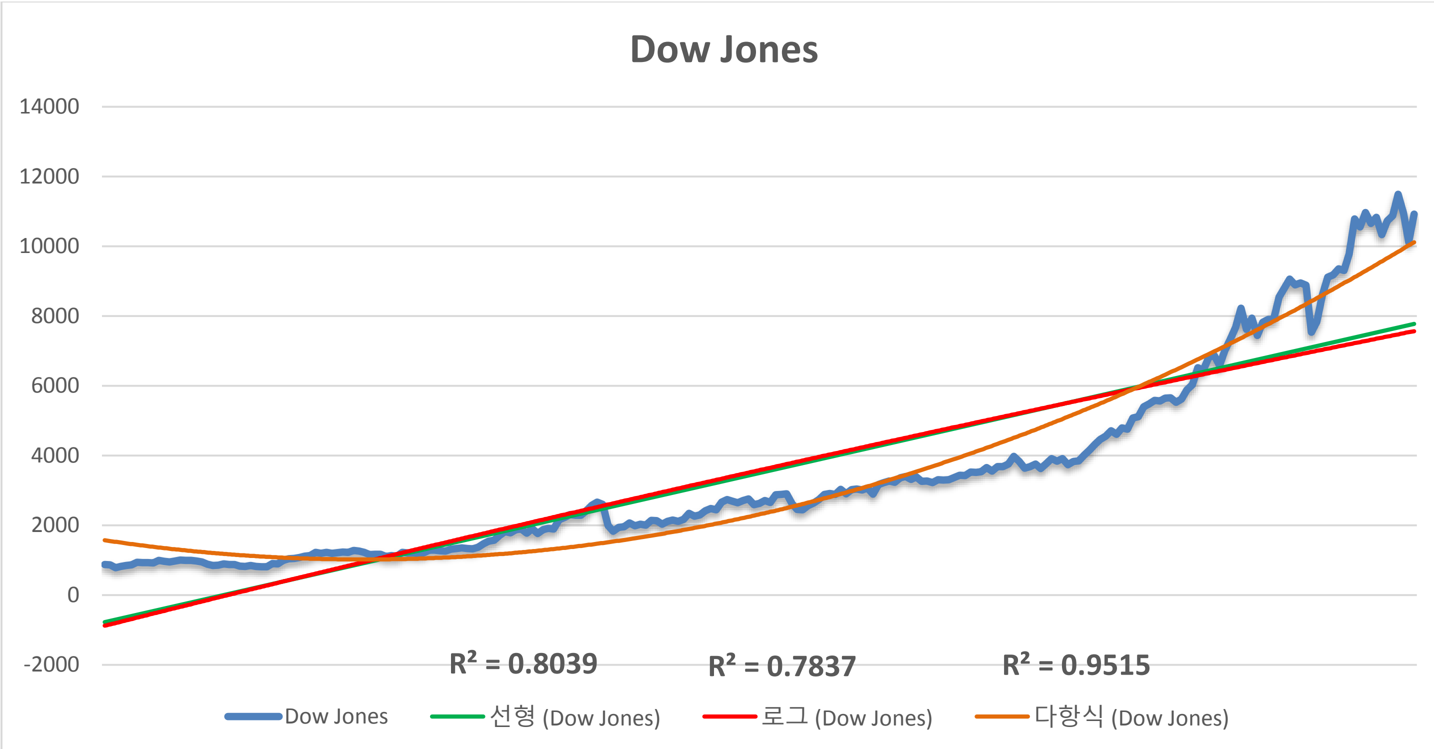
In the first analysis, we looked at the graphs as a whole and identified key portions of the data that seemed statistically promising or called for further, in-depth examination. Our search unearthed two distinct key portions; the sharp increase from 1980 to 2000, and the later sharp decrease in 2007-2008. In this analysis, we will be looking at the former feature, which also coincides with an important event in global economics; the so-called “dot-com bubble,”.

This phenomenon began when electronics and information technology rapidly developed from the 1980s to the 2000s, providing numerous companies, namely Apple, Microsoft, Intel, and others, with very rapid growth. This resulted in huge inflation of the stock market as a whole, as new technology began to increase exponentially in value. The “bubble” eventually burst in early 2000, when several of these information tech businesses (or “dot-coms”), suddenly failed. This led to a sharp drop in the stock market overall until most of the companies involved either rebounded or declared complete bankruptcy.

i. Dow Jones: 1980-2000

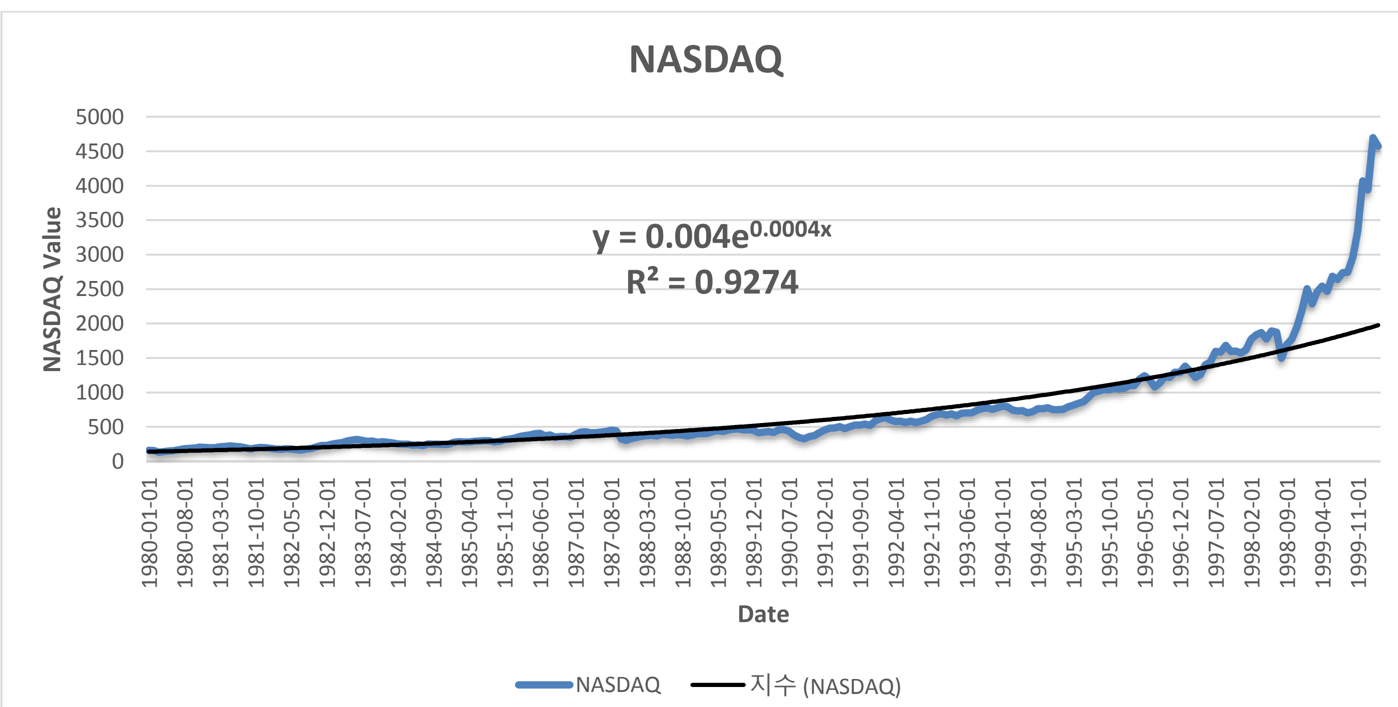
In this section, we will take only the points that represent the Dow from January 1980 to March 2000, during the dot-com bubble. Here, we see that the trend clearly is exponential, with the exponential factor being similar to that of the whole graphs, but with a much higher R^2 value at 0.9731, as the data is rid of outliers that were present in Analysis I.

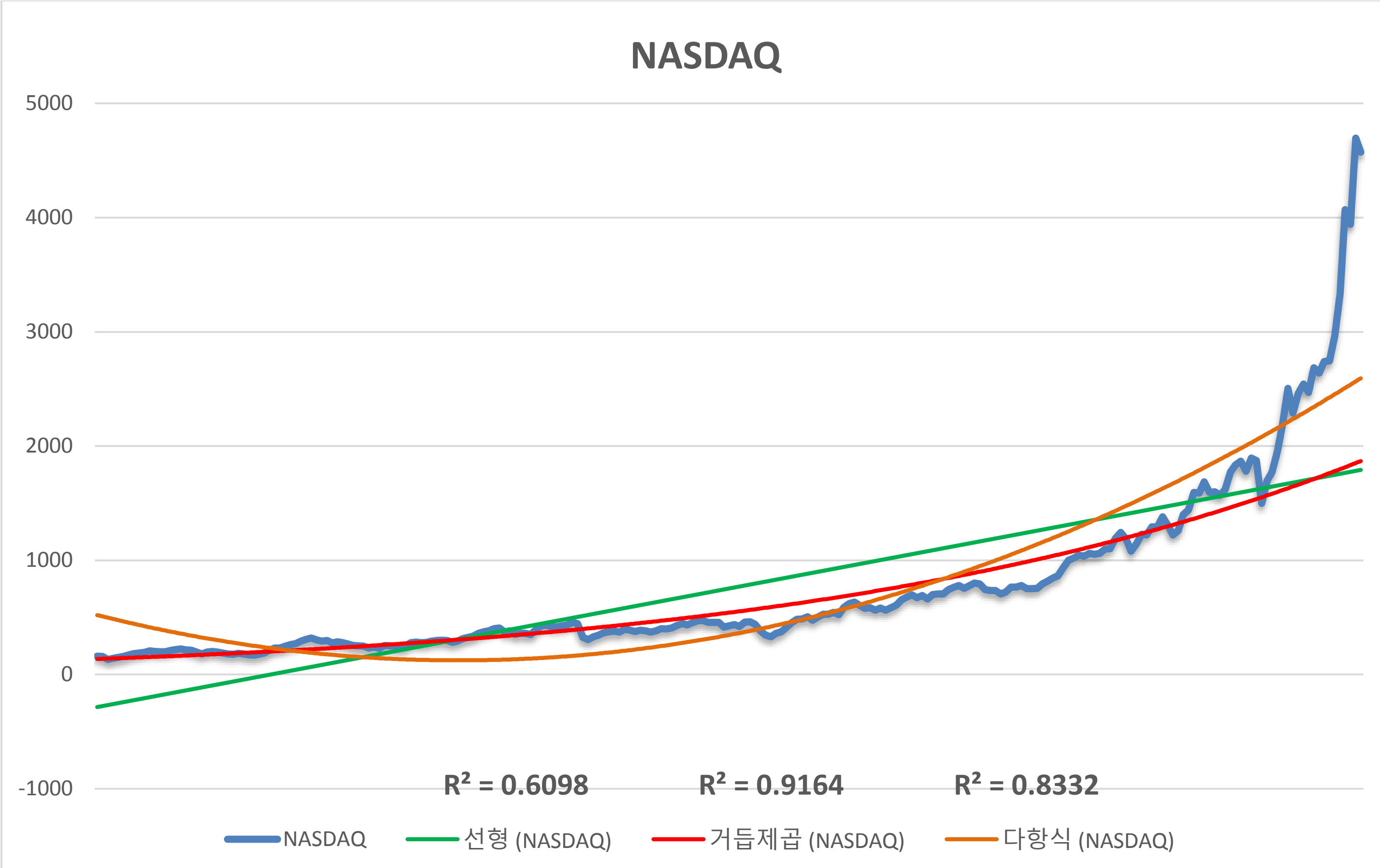




ii. NASDAQ: 1980-2000

In the NASDAQ, where the exponential increase was originally pointed out, although the visual appeal is much greater than in the Dow, the inconsistency of the increase (observe the sudden acceleration after January 1998, as shown by the green arrow) lessens the validity of the exponential trend. Nevertheless, the graph still shows a respectable R^2 value and an exponential factor not far from those in Analysis I.

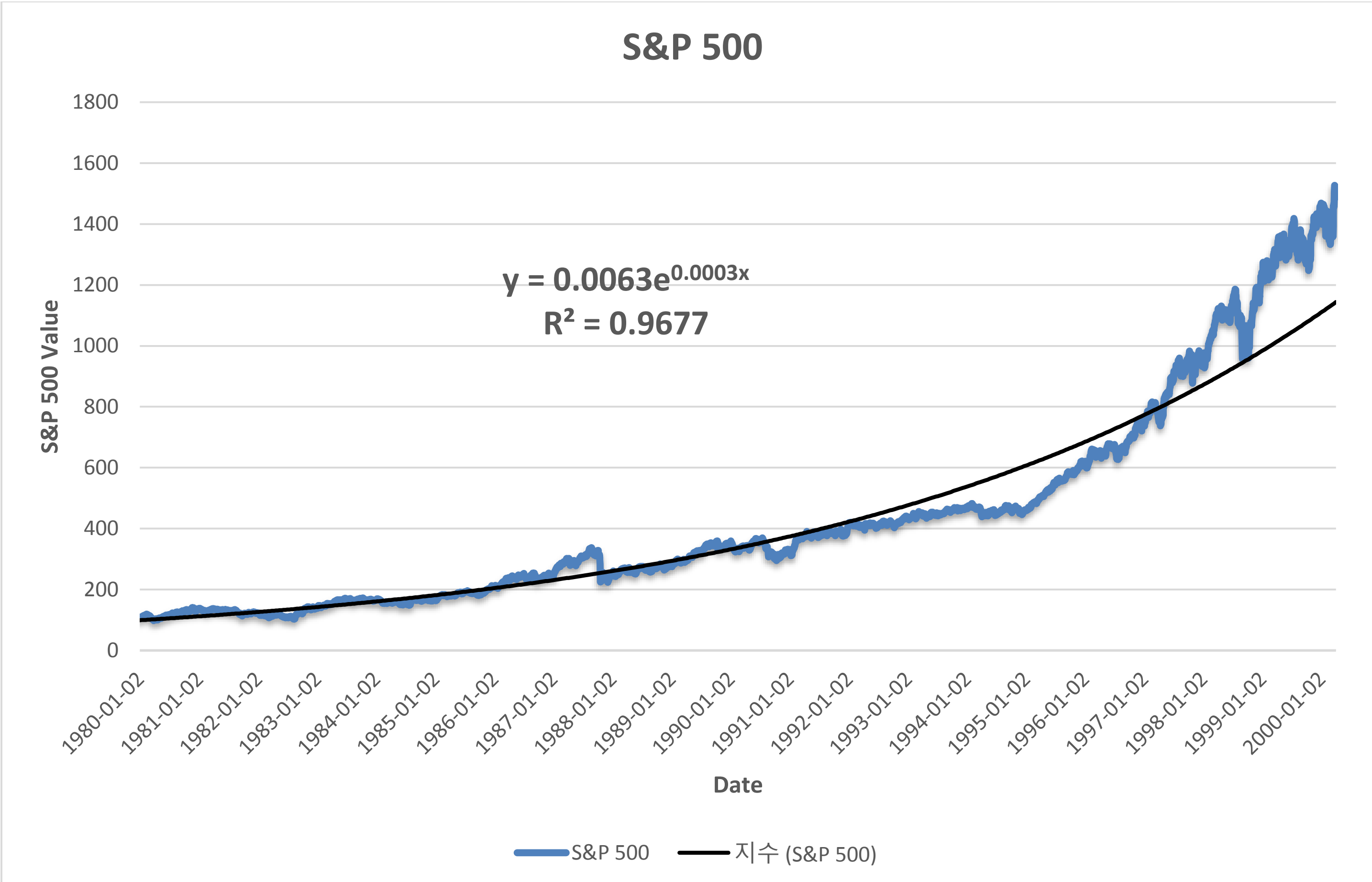


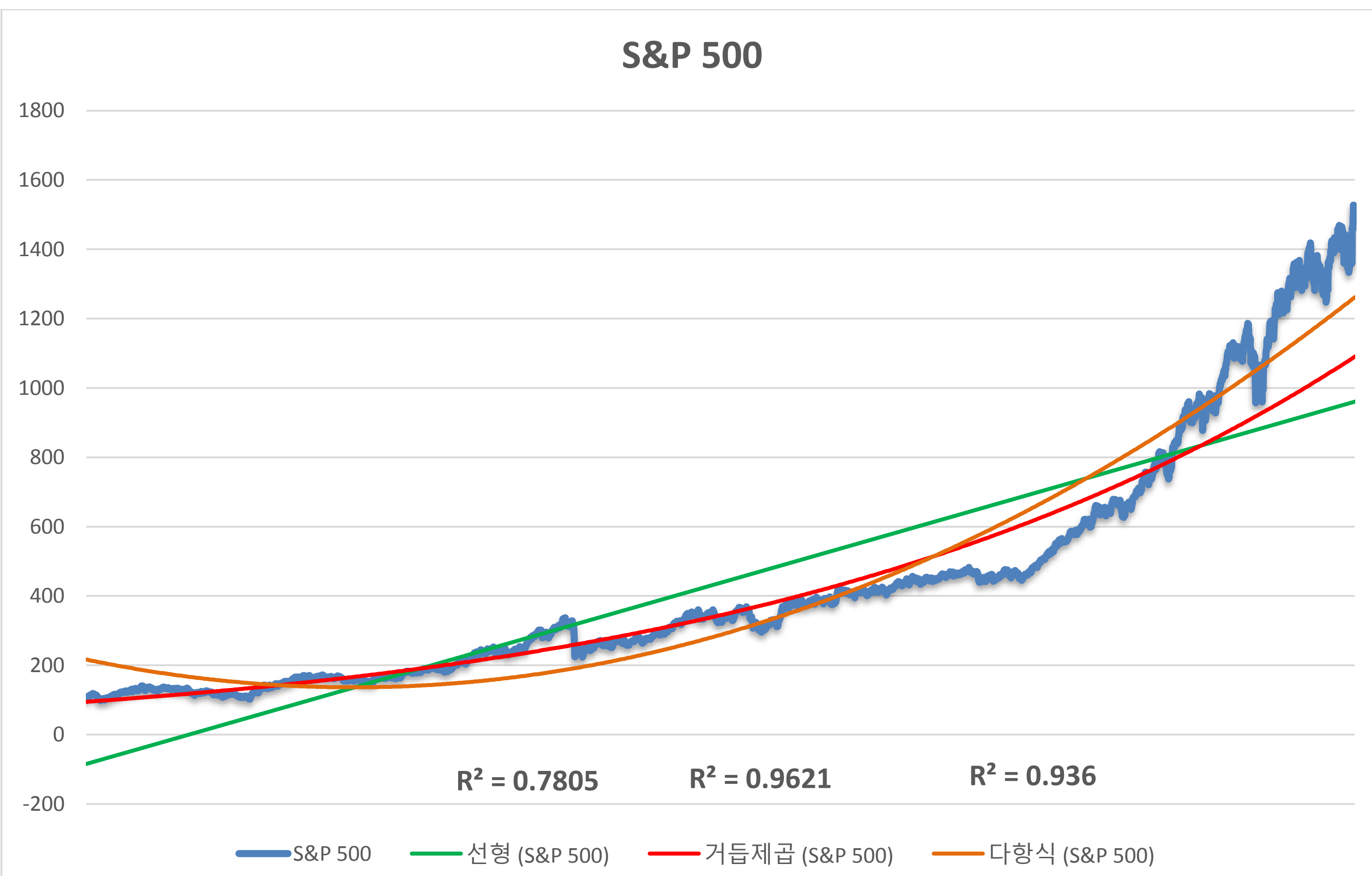


iii. S&P 500: 1980-2000

For the purposes of this analysis, I have introduced a new stock index, the Standard & Poor 500, which contains more diversified data from more companies than both the NASDAQ and the Dow Jones. Another useful point about the S&P 500 is that it publishes data sets for each day, not each month. This leads to a massive data set of more than 23,000 elements total, which is too broad for total analysis but fit for analyzing individual chunks, as here.

By taking the section from 1980 to 2000 – a full 5000 data points – and analyzing it, we can deduce a very clear exponential pattern in the data. Not only is the R^2 value extremely high at 0.9677, the exponential factor is also similar to that of all the other data sets examined during this lecture; all of them are somewhere between $e^{0.0002}$ and $e^{0.0004}$. This implies that all stock indexes increase at a reasonably constant rate per day, and therefore gives rise to the idea that the stock market itself increases at such a rate. Since despite their different calculation methods, stock indexes are all representations of the market as a whole, I deduce that this theory is plausible.

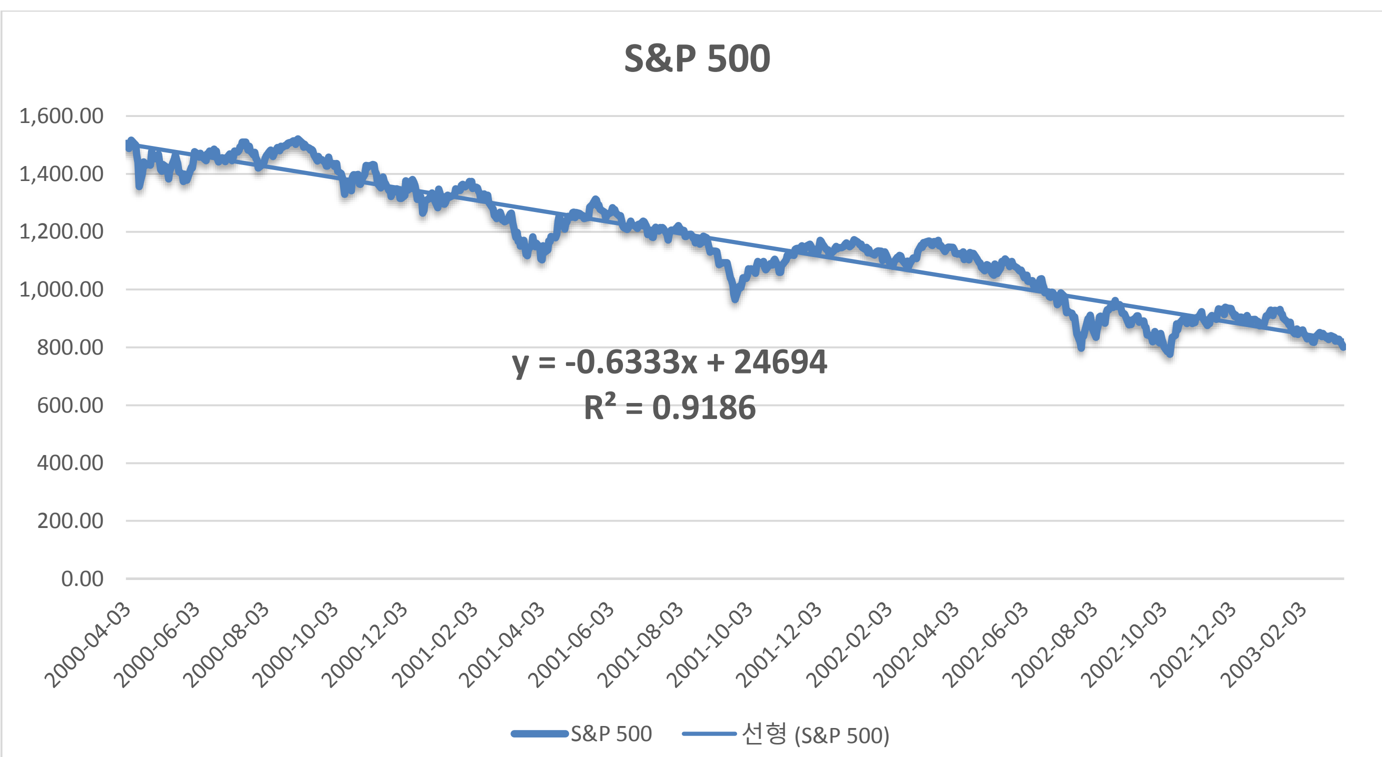

S&P 500
1800
1600
1400
1200
1000
800
600
400
200
0
S&P 500 Value
y = 0.0063e^0.0003x
R² = 0.9677
1980-01-02
1981-01-02
1982-01-02
1983-01-02
1984-01-02
1985-01-02
1986-01-02
1987-01-02
1988-01-02
1989-01-02
1990-01-02
1991-01-02
1992-01-02
1993-01-02
1994-01-02
1995-01-02
1996-01-02
1997-01-02
1998-01-02
1999-01-02
2000-01-02
Date
S&P 500
지수 (S&P 500)


S&P 500
1800
1600
1400
1200
1000
800
600
400
200
0
-200
R² = 0.7805
R² = 0.9621
R² = 0.936
S&P 500
선형 (S&P 500)
거듭제곱 (S&P 500)
다항식 (S&P 500)

iv. S&P 500: 2000-2003

Now, using the stock index newly introduced in Section III, we can evaluate smaller chunks of it; namely, the bursting of the "dot-com bubble" from March 2000 to March 2003. This would have previously been impossible due to the sparse data; a three-year data set would only contain thirty-six data points, far too less to draw a strong conclusion. However, thanks to the day-to-day nature of Standard & Poor 500, one can look into such short-term trends with confidence as well.

Surprisingly, the decline graph shows a linear trend, the first and last we will see in this lecture. Here, according to the x-coefficient, the Standard & Poor 500 decreased around -0.6333 points per day during the bursting of the dot-com bubble. One thing to also keep in mind here is that the y-intercept 24,694 is not the starting value; according to Excel's serial number system, 2000-04-03 is 100 years, 3 months (January, February, March), and 4 days ahead of 1900-01-01, which is 1. Furthermore, January and March always have 31 days each, and 2000 is **not** a leap year (Multiples of 100 are not leap years in the Gregorian calendar), so therefore, the actual starting point for the x-axis is $1 + 365 \times 100 + 31 + 29 + 31 + 4 = 36596$, and the starting value for y is $-0.6333 \times 36596 + 24694 \approx 1517.75$. Finally, according to this graph, if the trend was to continue, we can calculate the x-intercept; $x = 24694 \div 0.6333 \approx 38992.6$. Therefore, the S&P 500 would have dropped below 0 at x = 38993, or April 22, 2008.



5. Conclusion

Overall, in this lecture, we examined different types of correlations present in the stock market, and used them to analyze general movements of the data and point out specific regions where drastic change occurred. Later, in Analysis II, we focused on one such anomaly, which was also related to an event in economic history, the “dot-com bubble”. Here, we inspected both sides of this phenomenon; the astonishing exponential increase, and the sharp linear drop that followed. From that, we were able to identify the impact that information technology had on the stock market as a whole. Furthermore, we were able to make predictions about what could have happened otherwise using the data set, and also created a small theory concerning exponential growth factors in the stock market. Finally, I hope this was a fruitful time for all of you, and thank you for reading this paper.