DATA 333 Midterm

Thomas Soupionis

2025-03-25

1. Research Question

My research question: What is the historical probability of achieving a positive return when investing in major U.S. stock index funds (S&P 500, NASDAQ Composite, Dow Jones Industrial Average) over holding periods of 2, 5, and 10 years?

2. Data Source

I use daily closing price data for the S&P 500 (^SPX), NASDAQ Composite (^NDQ), and Dow Jones Industrial Average (^DJI) downloaded as CSV files from Stooq. The data is publically available at https://stooq.com/q/?s=%5Espx, https://stooq.com/q/?s=%5Endq, and https://stooq.com/q/?s=%5Edji. The unit of analysis is the investment start date (each trading day).

3. Importing Data

I run the following R code to import the data:

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.3
```

```
## Warning: package 'forcats' was built under R version 4.3.3
```

```
## Warning: package 'lubridate' was built under R version 4.3.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)

dow_data <- read_csv("../data/official_dowjones_data.csv")
```

```
## Rows: 33336 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Date
## dbl (1): Close
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
spx_data <- read_csv("../data/official_spx_data.csv")
```

```
## Rows: 17116 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Date
## dbl (1): Close
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
nasdaq_data <- read_csv("../data/official_nasdaq_data.csv")
```

```
## Rows: 13635 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Date
## dbl (1): Close
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Here is a look at a few records of the imported data:

```
head(dow_data)
```

```
## # A tibble: 6 x 2
##   Date       Close
##   <chr>      <dbl>
## 1 1896-05-27  29.4
## 2 1896-05-28  29.1
## 3 1896-05-29  29.4
## 4 1896-06-01  29.4
## 5 1896-06-02  29
## 6 1896-06-03  28.8
```

```r
head(spx_data)
```

```
## # A tibble: 6 x 2
##   Date      Close
##   <chr>     <dbl>
## 1 3/4/1957   44.1
## 2 3/5/1957   44.2
## 3 3/6/1957   44.2
## 4 3/7/1957   44.2
## 5 3/8/1957   44.1
## 6 3/11/1957  43.8
```

```r
head(nasdaq_data)
```

```
## # A tibble: 6 x 2
##   Date      Close
##   <chr>     <dbl>
## 1 2/5/1971   100
## 2 2/8/1971   101.
## 3 2/9/1971   101.
## 4 2/10/1971  101.
## 5 2/11/1971  102.
## 6 2/12/1971  102.
```

4. Creating the Dependent Variable

The dependent variable is the percentage return of each index over specified holding periods, calculated as (Closing_Price_end / Closing_Price_start - 1) * 100

```r
calculate_returns <- function(data) {
  data %>% # using tidyverse pipe operator
    rename(Exit_Date = Date,
           Closing_Price = Close
    ) %>%
    mutate(
      Exit_Date = as.character(Exit_Date), # Ensure character format
      Exit_Date = parse_date_time(Exit_Date, orders = c("Y-m-d", "m/d/Y"))) %>%
    # parse_date_time automatically handles both YYYY-MM-DD and M/D/YYYY format
    arrange(Exit_Date) %>%
    # sort rows chronologically so later calculations use past prices in the right order
    mutate(
      Return_2yr  = (Closing_Price / lag(Closing_Price, 504) - 1) * 100,
      Return_5yr  = (Closing_Price / lag(Closing_Price, 1260) - 1) * 100,
      Return_10yr = (Closing_Price / lag(Closing_Price, 2520) - 1) * 100
    ) # getting percentage returns for each holding period
  # lag allows you to compare each value to some future value
  # Ex: 504 is about 2 years later in trading days, 1260 is about 5 years later in trading days, etc.)
}

dow_returns   <- calculate_returns(dow_data)
spx_returns   <- calculate_returns(spx_data)
```

```r
nasdaq_returns <- calculate_returns(nasdaq_data)
# 3 new columns added for each table: Return_2yr, Return_5yr and Return_10yr
```

5. Creating the Independent Variable

The independent variable is the holding period (2, 5, 10 years), represented by the return columns (Return_2yr, Return_5yr, Return_10yr) -> in getting the dependent variable above I also created the independent variables

6. Other Necessary Data Processing

```r
all_returns <- bind_rows(
  dow_returns   %>% mutate(Index = "Dow Jones"),
  spx_returns   %>% mutate(Index = "S&P 500"),
  nasdaq_returns %>% mutate(Index = "NASDAQ")
  # Add an "Index" column to distinguish which dataset each row comes from
) %>%
  pivot_longer(
    cols = starts_with("Return"), # select all return columns
    names_to = "Horizon", # new column to store the return period
    values_to = "Return" # new column to store the corresponding return values
  ) %>%
  # Convert wide format (separate columns for each return period) to long format
  mutate(Horizon = case_when(
    Horizon == "Return_2yr"  ~ "2 Years",
    Horizon == "Return_5yr"  ~ "5 Years",
    Horizon == "Return_10yr" ~ "10 Years"
  )) %>%
    # renames return columns for clarity
  filter(!is.na(Return))
  # removes rows where the Return is NA (not enough past data)
```

This analysis examines the historical probability of achieving a positive return when investing in major U.S. stock indices over 2-, 5-, and 10-year holding periods. By leveraging daily closing prices for the Dow Jones, S&P 500, and NASDAQ, returns were calculated based on historical price movements. The results indicate that returns generally become more stable and positive over longer investment horizons, reinforcing the principle that long-term investing reduces the risk of negative outcomes. While minor discrepancies exist due to the assumption that 504, 1260, and 2520 trading days correspond exactly to 2, 5, and 10 years, the methodology remains a reasonable approximation for assessing historical market trends. These findings support the broader financial argument that index investing over extended periods historically yields favorable returns.