Forecasting Constituents of the MSCI Minimum Volatility Index Through Logistic Regression

---

A Thesis
Presented to
The Division of Statistics
Harvard College

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts in Statistics (Honors)

John A. Gilheany

November 6, 2017

Approved for the Division
(Statistics)

Professor Michael Parzen	David Kane

Acknowledgements

I want to thank Prof. Parzen and David Kane for all of their help.

Preface

This thesis explores a way of predicting index constituents using logistic regression.

Table of Contents

List of Tables

List of Figures

Abstract

The low-risk anomaly has created opportunities for arbitrage in the financial markets. As Baker et al. discuss in "Benchmarks as Limits to Arbitrage: Understanding the Low-Volatility Anomaly," low-volatility and low-beta portfolios outperform and high-volatility and high-beta portfolios by a factor of several times due to benchmarking and lottery-preferences. The iShares MSCI USA Minimum Volatility (USMV) is an ETF tracking a minimum volatility index that was used to find data and will be used for trading arbitrage. Frazzini et al. discuss arbitrage opportunities by quantitative focused funds like AQR in "Betting Against Beta", and this thesis explores a more advanced type of index front-running as a potential arbitrage opportunity. Data was collected from USMV from its inception in October 2011, and from EUSA, the parent ETF of USMV, from the same period until December 2016. 52-week trailing beta, 52-week trailing volatility, lagged price/book, and current index membership were calculated, and a regression model was run to quantify the relationship between current index membership and these four variables. In the model, a probabilities of index membership were calculated and an optimal cutoff was calculated to which the model would be 95% accurate of its findings of a stock to be in or out of USMV, given the historical data. Backtesting with prior data showed with a model accuracy of 95%, arbitrage opportunities of X% could be collected after each rebalancing.

Chapter 1

Introduction

1.1 Background

The iShares USA Minimum Volatility (USMV) Exchange Traded Fund (ETF) is designed to track the investment results of the MSCI Minimum Volatility USA index, which is composed of stocks with a lower volatility than the general market. This can provide investors with exposure to a portfolio with less risk than many alternatives, and historically has declined less in value than the broader market during economic downturns. The ETF is comprised of 189 holdings, and is rebalanced two times per year. The purpose of this dissertation is to create a logistic regression model that can accurately predict which stocks will be added or removed from this ETF before rebalancing occurs, and understand what factors are involved. The model will take into account volatility attributes of each stock, as well as others potentially significant predictor variables from prior studies. An accurate model will allow for arbitrage investment opportunities.

1.1.1 Exchange Traded Funds

An Exchange Traded Fund (ETF) is a collection of stocks and/or bonds in a single portfolio, that is traded on a major exchange just like a stock is (Hayes, 2017). As a result, the price of an ETF fluctuates on a regular basis. Exchange Traded Funds generally have more liquidity and less fees when compared to other alternatives instruments like mutual funds. Owning an ETF can allow investors to minimize risk, since owning an ETF is comparable to owning a little bit of many different stocks. This diversification comes at lower costs and less effort for investors as well.

ETFs can also track an index, commodity, bonds, or basket of all of the above. Unlike an ETF, which is publicly traded, an index is not. The goal of the USMV ETF is to track the MSCI Minimum Volatility USA index, and this is more complicated than it seems. In addition to tracking this index, the ETF aims to mirror returns of the index and any difference is called tracking error. Many times, the tracking error is often very small, and can be around a tenth of a percent. This error can come from indices being market capitalization weighted, meaning that for each price fluctuations of each stock lead to the weighting being changed by a ratio of its market cap against

the market cap of all stocks in the index (http://www.investopedia.com/articles/exchangetradedfunds/09/tracking-error-etf-funds.asp). With these stocks weightings in the index constantly changing and people buying in and out of ETFs constantly, it is hard to track performance entirely accurately. However, ETFs very closely follow indices, as their tracking errors are generally quite small. Thus, although ETF data is not the same as index data, they are very similar.

1.1.2 iShares MSCI Min Vol USA ETF

The iShares MSCI Min Vol USA ETF (USMV) is a Blackrock-managed ETF that tracks the investment results of the MSCI Minimum Volatility USA index. The MSCI Minimum Volatility USA index constituents come from the MSCI USA Index, which are roughly comprised of the top 600 US stocks by market cap. This minimum volatility index is intended to have a lower beta, lower volatility, lower cap bias, and contain more stocks with less risk than its parent index, which contains US mid-cap and large-cap stocks. The index is rebalanced twice a year, on the last trading days of May and November. The index typically has around 180 constituents, with an average of 20 new additions and 14 deletions every 6 months when rebalancing occurs. Over the last five, years, the number of additions has ranged from 12 to 25, while the deletions have been between 10 and 19. Changes to the index are usually announced nine trading days before they are set to take place.

Using the Barra Open Optimizer, USMV creates a minimum variance portfolio of low risk stocks, as a subset from its parent index of USA large-cap and mid-cap stock. Using this estimated security covariance matrix, the MSCI Minimum Volatility Index is the product of the lowest absolute volatility, considering the constraints. Moreover, these additions are simply a relabeling of existing stocks in the parent index, and do not include new additions to the parent index. The low-risk stocks chosen to be in USMV are determined by a set of constraints, like maintaining a certain sector or country weight relative to the parent index.

There are many specific constraints to this index. The first is that an individual stock cannot exceed 1.5% or 20 times the weight of the stock in the parent index. The minimum weight of a security in the index is also capped at 0.05%. USMV also aims to keep the weight of specific countries within a 5% range of the weight in the parent index, or 3 times the weight of the country in the parent index. Sector weights of USMV also cannot deviate more than 5% from the sector weights in the parent index. One way turnover of the index is also maxed at 10%. Thus, taking into account these constraints, the Barra Open Optimizer creates the lowest absolute volatility portfolio possible (https://seekingalpha.com/article/3964639-understanding-ishares-msci-usa-minimum-volatility-etf)

1.1.3 Purpose

As mentioned, the purpose of this thesis is to create a model to that will predict rebalancing of stocks in the Min Vol index, and thus the USMV ETF, before it actually happens. There is significant price movement whenever a stock is added or removed

from a large ETF, like USMV. When a stock is added to the index, the ETF will buy large amounts of that stock, increasing the demand, and consequently market price for that stock. If the stock is bought in advance of this large purchase, then the investor can enjoy pretty immediate price appreciation in the stock. Moreover, if a stock is removed from the Min Vol index, the USMV ETF will sell all current holdings of the stock, which would increase the supply of the stock, driving down market price of the stock. If one were to short this stock before that happened, he/she can also profit from that event.

A phenomena known as ETF front-running has been around for a long time and is similar to what this paper hopes to accomplish, but is one step behind. ETF front-running involves traders buying or selling stocks in advance of ETF managers after they announce an exit or entrance of a position (`https://seekingalpha.com/article/165877-how-traders-are-front-running-etfs`). There is typically a slight lag between an announcement of an ETF to add or remove a position, and the actual purchase or sale of this position. By acting quickly, traders can scalp profit by buying a stock before an ETF does, and selling it to them at a slight profit, or short-selling a stock before an ETF exits the position, and then buying it back at the lower price. The thesis will take this one step farther, and try predict the stock addition or deletion before announcement. This will allow traders to similar front-run the index, but they will do so before the market is able to react, leading to larger profit opportunities.

1.1.4 Logistic Regression Model

These goals of this paper will be achieved by creating a logistic regression model, which will be transformed to calculate a probability of a stock being in the out of the index. The predictor variables will include 52-week trailing volatility, 52-week trailing beta, price/book ratio, and whether or not the stock was in the index 6 during the previous rebalancing. These attributes were chosen after looking at the historical literature and understanding of the minimum volatility index.

Chapter 2

Mathematics and Science

2.1 Math

TEX is the best way to typeset mathematics. Donald Knuth designed TEX when he got frustrated at how long it was taking the typesetters to finish his book, which contained a lot of mathematics. One nice feature of *R Markdown* is its ability to read LaTeX code directly.

If you are doing a thesis that will involve lots of math, you will want to read the following section which has been commented out. If you're not going to use math, skip over or delete this next commented section.

2.2 Chemistry 101: Symbols

Chemical formulas will look best if they are not italicized. Get around math mode's automatic italicizing in LaTeX by using the argument `$\mathrm{formula here}$`, with your formula inside the curly brackets. (Notice the use of the backticks here which enclose text that acts as code.)

So, $Fe_2^{2+}Cr_2O_4$ is written `$\mathrm{Fe_2^{2+}Cr_2O_4}$`.
Exponent or Superscript: O^-
Subscript: CH_4

To stack numbers or letters as in Fe_2^{2+}, the subscript is defined first, and then the superscript is defined.
Bullet: $CuCl \bullet 7H_2O$
Delta: Δ
Reaction Arrows: $\longrightarrow$ or $\xrightarrow{solution}$
Resonance Arrows: $\leftrightarrow$
Reversible Reaction Arrows: $\rightleftharpoons$

2.2.1 Typesetting reactions

You may wish to put your reaction in an equation environment, which means that LaTeX will place the reaction where it fits and will number the equations for you.

$$\mathrm{C_6H_{12}O_6 + 6O_2 \longrightarrow 6CO_2 + 6H_2O} \tag{2.1}$$

We can reference this combustion of glucose reaction via Equation (2.1).

2.2.2 Other examples of reactions

$\mathrm{NH_4Cl_{(s)} \rightleftharpoons NH_{3(g)} + HCl_{(g)}}$
$\mathrm{MeCH_2Br + Mg \xrightarrow[below]{above} MeCH_2 \bullet Mg \bullet Br}$

2.3 Physics

Many of the symbols you will need can be found on the math page `http://web.reed.edu/cis/help/latex/math.html` and the Comprehensive LaTeX Symbol Guide (`http://mirror.utexas.edu/ctan/info/symbols/comprehensive/symbols-letter.pdf`).

2.4 Biology

You will probably find the resources at `http://www.lecb.ncifcrf.gov/~toms/latex.html` helpful, particularly the links to bsts for various journals. You may also be interested in TeXShade for nucleotide typesetting (`http://homepages.uni-tuebingen.de/beitz/txe.html`). Be sure to read the proceeding chapter on graphics and tables.

Chapter 3

Tables, Graphics, References, and Labels

3.1 Tables

In addition to the tables that can be automatically generated from a data frame in **R** that you saw in [R Markdown Basics] using the `kable` function, you can also create tables using *pandoc*. (More information is available at `http://pandoc.org/README.html#tables`.) This might be useful if you don't have values specifically stored in **R**, but you'd like to display them in table form. Below is an example. Pay careful attention to the alignment in the table and hyphens to create the rows and columns.

Table 3.1: Correlation of Inheritance Factors for Parents and Child

Factors	Correlation between Parents & Child	Inherited
Education	-0.49	Yes
Socio-Economic Status	0.28	Slight
Income	0.08	No
Family Size	0.18	Slight
Occupational Prestige	0.21	Slight

We can also create a link to the table by doing the following: Table 3.1. If you go back to [Loading and exploring data] and look at the `kable` table, we can create a reference to this max delays table too: Table **??**. The addition of the `(\#tab:inher)` option to the end of the table caption allows us to then make a reference to Table `\@ref(tab:label)`. Note that this reference could appear anywhere throughout the document after the table has appeared.

3.2 Figures

If your thesis has a lot of figures, *R Markdown* might behave better for you than that other word processor. One perk is that it will automatically number the figures accordingly in each chapter. You'll also be able to create a label for each figure, add a caption, and then reference the figure in a way similar to what we saw with tables earlier. If you label your figures, you can move the figures around and *R Markdown* will automatically adjust the numbering for you. No need for you to remember! So that you don't have to get too far into LaTeX to do this, a couple **R** functions have been created for you to assist. You'll see their use below.

In the **R** chunk below, we will load in a picture stored as `reed.jpg` in our main directory. We then give it the caption of "Reed logo", the label of "reedlogo", and specify that this is a figure. Make note of the different **R** chunk options that are given in the R Markdown file (not shown in the knitted document).

```
include_graphics(path = "figure/reed.jpg")
```

Figure 3.1: Reed logo

Here is a reference to the Reed logo: Figure 3.1. Note the use of the `fig:` code here. By naming the **R** chunk that contains the figure, we can then reference that figure later as done in the first sentence here. We can also specify the caption for the figure via the R chunk option `fig.cap`.

Below we will investigate how to save the output of an **R** plot and label it in a way similar to that done above. Recall the `flights` dataset from Chapter **??**. (Note that we've shown a different way to reference a section or chapter here.) We will next explore a bar graph with the mean flight departure delays by airline from Portland for 2014. Note also the use of the `scale` parameter which is discussed on the next page.

```
flights %>% group_by(carrier) %>%
  summarize(mean_dep_delay = mean(dep_delay)) %>%
  ggplot(aes(x = carrier, y = mean_dep_delay)) +
  geom_bar(position = "identity", stat = "identity", fill = "red")
```

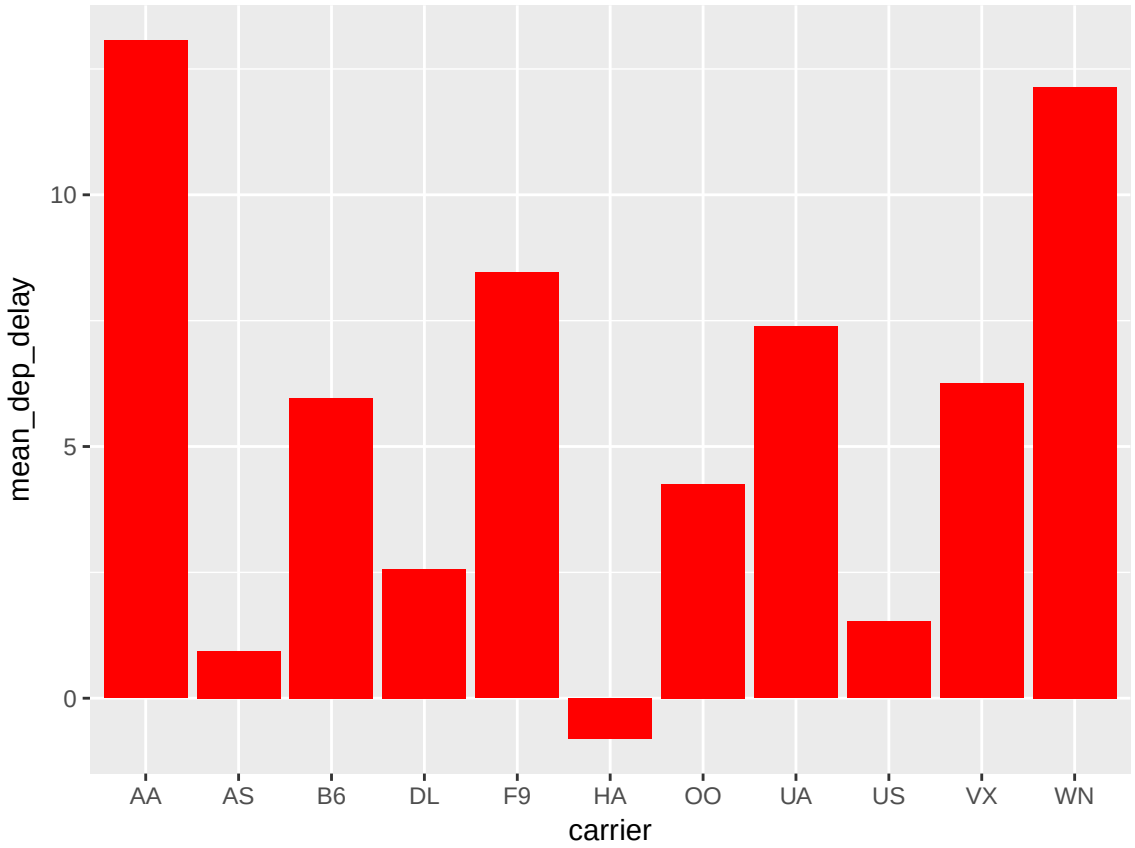

Figure 3.2: Mean Delays by Airline

Here is a reference to this image: Figure 3.2.

A table linking these carrier codes to airline names is available at `https://github.com/ismayc/pnwflights14/blob/master/data/airlines.csv`.

Next, we will explore the use of the `out.extra` chunk option, which can be used to shrink or expand an image loaded from a file by specifying `"scale= "`. Here we use the mathematical graph stored in the "subdivision.pdf" file.

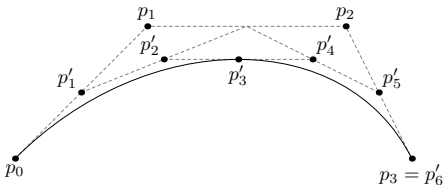

Figure 3.3: Subdiv. graph

Here is a reference to this image: Figure 3.3. Note that `echo=FALSE` is specified so that the **R** code is hidden in the document.

More Figure Stuff

Lastly, we will explore how to rotate and enlarge figures using the `out.extra` chunk option. (Currently this only works in the PDF version of the book.)

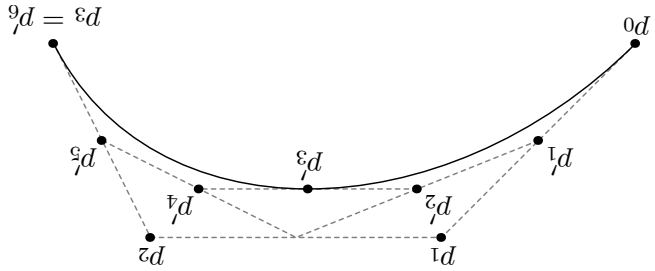

Figure 3.4: A Larger Figure, Flipped Upside Down

As another example, here is a reference: Figure 3.4.

3.3 Footnotes and Endnotes

You might want to footnote something.[1] The footnote will be in a smaller font and placed appropriately. Endnotes work in much the same way. More information can be found about both on the CUS site or feel free to reach out to `data@reed.edu`.

3.4 Bibliographies

Of course you will need to cite things, and you will probably accumulate an armful of sources. There are a variety of tools available for creating a bibliography database (stored with the .bib extension). In addition to BibTeX suggested below, you may want to consider using the free and easy-to-use tool called Zotero. The Reed librarians have created Zotero documentation at `http://libguides.reed.edu/`

[1] footnote text

`citation/zotero`. In addition, a tutorial is available from Middlebury College at `http://sites.middlebury.edu/zoteromiddlebury/`.

R Markdown uses *pandoc* (`http://pandoc.org/`) to build its bibliographies. One nice caveat of this is that you won't have to do a second compile to load in references as standard LaTeX requires. To cite references in your thesis (after creating your bibliography database), place the reference name inside square brackets and precede it by the "at" symbol. For example, here's a reference to a book about worrying: (Molina & Borkovec, 1994). This `Molina1994` entry appears in a file called `thesis.bib` in the `bib` folder. This bibliography database file was created by a program called BibTeX. You can call this file something else if you like (look at the YAML header in the main .Rmd file) and, by default, is to placed in the `bib` folder.

For more information about BibTeX and bibliographies, see our CUS site (`http://web.reed.edu/cis/help/latex/index.html`)[2]. There are three pages on this topic: *bibtex* (which talks about using BibTeX, at `http://web.reed.edu/cis/help/latex/bibtex.html`), *bibtexstyles* (about how to find and use the bibliography style that best suits your needs, at `http://web.reed.edu/cis/help/latex/bibtexstyles.html`) and *bibman* (which covers how to make and maintain a bibliography by hand, without BibTeX, at `http://web.reed.edu/cis/help/latex/bibman.html`). The last page will not be useful unless you have only a few sources.

If you look at the YAML header at the top of the main .Rmd file you can see that we can specify the style of the bibliography by referencing the appropriate csl file. You can download a variety of different style files at `https://www.zotero.org/styles`. Make sure to download the file into the csl folder.

Tips for Bibliographies

- Like with thesis formatting, the sooner you start compiling your bibliography for something as large as thesis, the better. Typing in source after source is mind-numbing enough; do you really want to do it for hours on end in late April? Think of it as procrastination.
- The cite key (a citation's label) needs to be unique from the other entries.
- When you have more than one author or editor, you need to separate each author's name by the word "and" e.g. `Author = {Noble, Sam and Youngberg, Jessica},`.
- Bibliographies made using BibTeX (whether manually or using a manager) accept LaTeX markup, so you can italicize and add symbols as necessary.
- To force capitalization in an article title or where all lowercase is generally used, bracket the capital letter in curly braces.
- You can add a Reed Thesis citation[3] option. The best way to do this is to use the phdthesis type of citation, and use the optional "type" field to enter "Reed thesis" or "Undergraduate thesis."

[2] Reed College (2007)
[3] Noble (2002)

3.5 Anything else?

If you'd like to see examples of other things in this template, please contact the Data @ Reed team (email `data@reed.edu`) with your suggestions. We love to see people using *R Markdown* for their theses, and are happy to help.

Conclusion

If we don't want Conclusion to have a chapter number next to it, we can add the `{-}` attribute.

More info

And here's some other random info: the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.

Appendix A

The First Appendix

This first appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readibility and/or setup.

In the main Rmd file

In Chapter 3:

```
# This chunk ensures that the thesisdown package is
# installed and loaded. This thesisdown package includes
# the template files for the thesis and also two functions
# used for labeling and referencing
if(!require(devtools))
  install.packages("devtools", repos = "http://cran.rstudio.com")
if(!require(dplyr))
    install.packages("dplyr", repos = "http://cran.rstudio.com")
if(!require(ggplot2))
    install.packages("ggplot2", repos = "http://cran.rstudio.com")
if(!require(ggplot2))
    install.packages("bookdown", repos = "http://cran.rstudio.com")
if(!require(thesisdown)){
  library(devtools)
  devtools::install_github("ismayc/thesisdown")
  }
library(thesisdown)
flights <- read.csv("data/flights.csv")
```

Appendix B

The Second Appendix, for Fun

References

Angel, E. (2000). *Interactive computer graphics : A top-down approach with opengl*. Boston, MA: Addison Wesley Longman.

Angel, E. (2001a). *Batch-file computer graphics : A bottom-up approach with quicktime*. Boston, MA: Wesley Addison Longman.

Angel, E. (2001b). *Test second book by angel*. Boston, MA: Wesley Addison Longman.

Hayes, A. (2017, October). Exchange-traded fund (etf). Retrieved from `http://www.investopedia.com/terms/e/etf.asp`

Molina, S. T., & Borkovec, T. D. (1994). The Penn State worry questionnaire: Psychometric properties and associated characteristics. In G. C. L. Davey & F. Tallis (Eds.), *Worrying: Perspectives on theory, assessment and treatment* (pp. 265–283). New York: Wiley.

Noble, S. G. (2002). *Turning images into simple line-art* (Undergraduate thesis). Reed College.

Reed College. (2007, march). LaTeX your document. Retrieved from `http://web.reed.edu/cis/help/LaTeX/index.html`