**Boston University Questrom School of Business**

**MF 793 – Fall 2020**

Eric Jacquier

**Problem Set 1**

**Due Sunday September 27th at 10pm Boston Time**

Problems turned in after the deadline are not graded

* Do the Problem Set in groups of four at the most – students can be in different sections.
* Turn in one pdf copy for the group on the course site. No email, no paper submission.
* The best way is to start a word file which allows you to pick tables from this file
* **To get a check, you need to answer all the questions.**

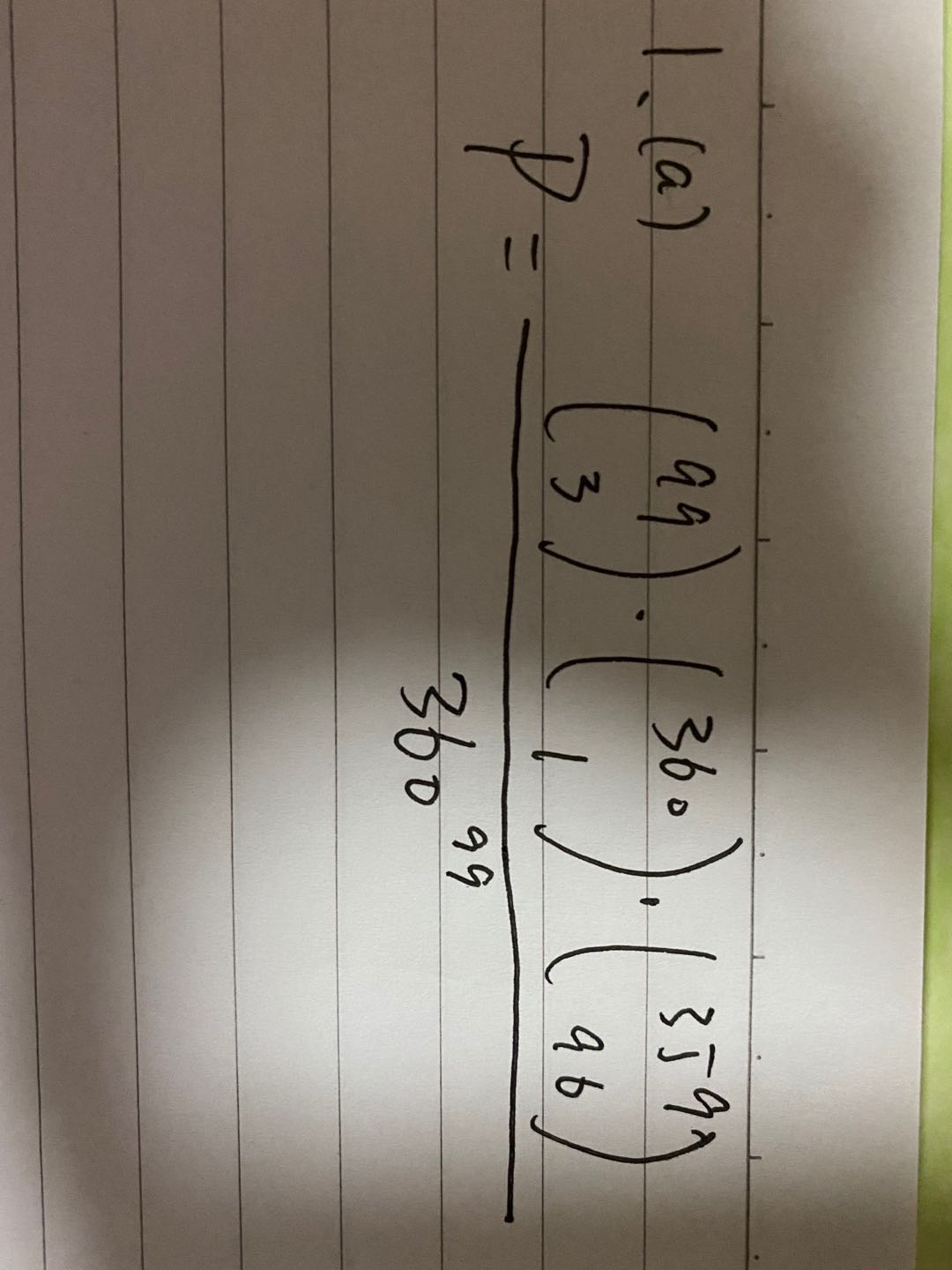
**If you do not do this, you can not get a check plus**

* ALL discussion and math questions answered.
* All math questions hand-written
* All Figures professionally made with X and Y axes labels and title and fig. numbers
* Tables must have row and column names, title and table number.
* Numbers in the tables must **not** contain too many useless or irrelevant digit, use your common sense as to how many digits to report in a Table. Otherwise it looks like you have no idea what matters.
* All R code as an appendix must be at the back of the homework.

**Problem 1:** Counting

The new middle school class has arrived, all 99 of them. The social events director wants to know the probability that exactly 3 kids (no more, no less) have their birthday the same day, and the rest all have different birthdays. Use a 360 day year.

1. Write by hand the theoretical formula.



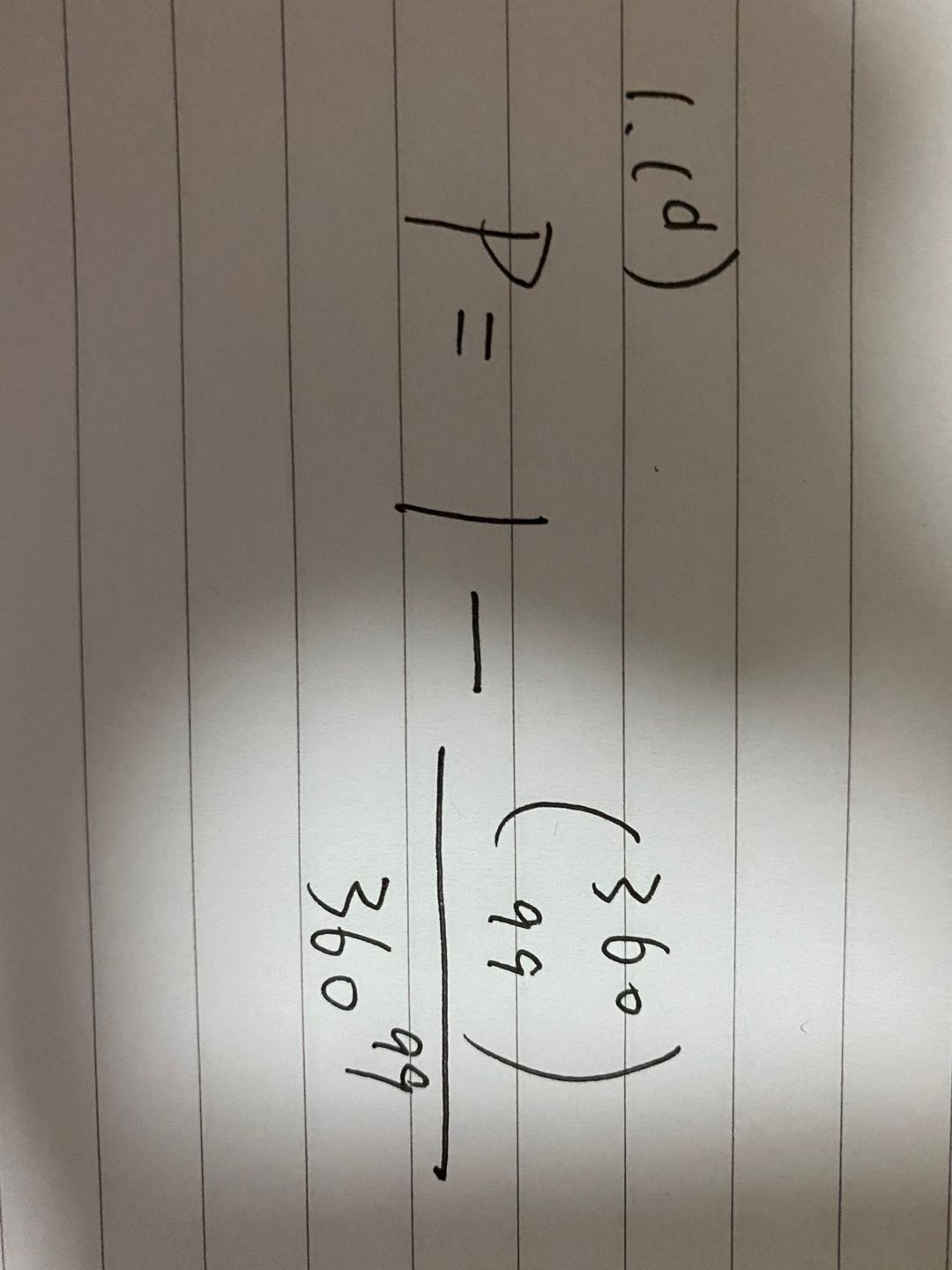
b) Explain in words how you came up with the numerator, and the denominator.

The numerator part means randomly choose 3 students that having the same birthday, the numerator part means randomly choose 1 day as the same birthday previously chosen 3 students share, the numerator part means randomly choose 96 days from other 359 days to allocate the birthday of other 96 students who have different birthdays. The denominator means total number of different situations of 99 students’ birthday. (each student can have a choice from 360 days as his/hers birthday randomly.)

c) Compute it in R and give the result. Don’t forget to show the code you used.

Probability = 7.71e-157

d) What is the probability that at least (possibly more) 2 students share a birthday? Give the formula and the final result



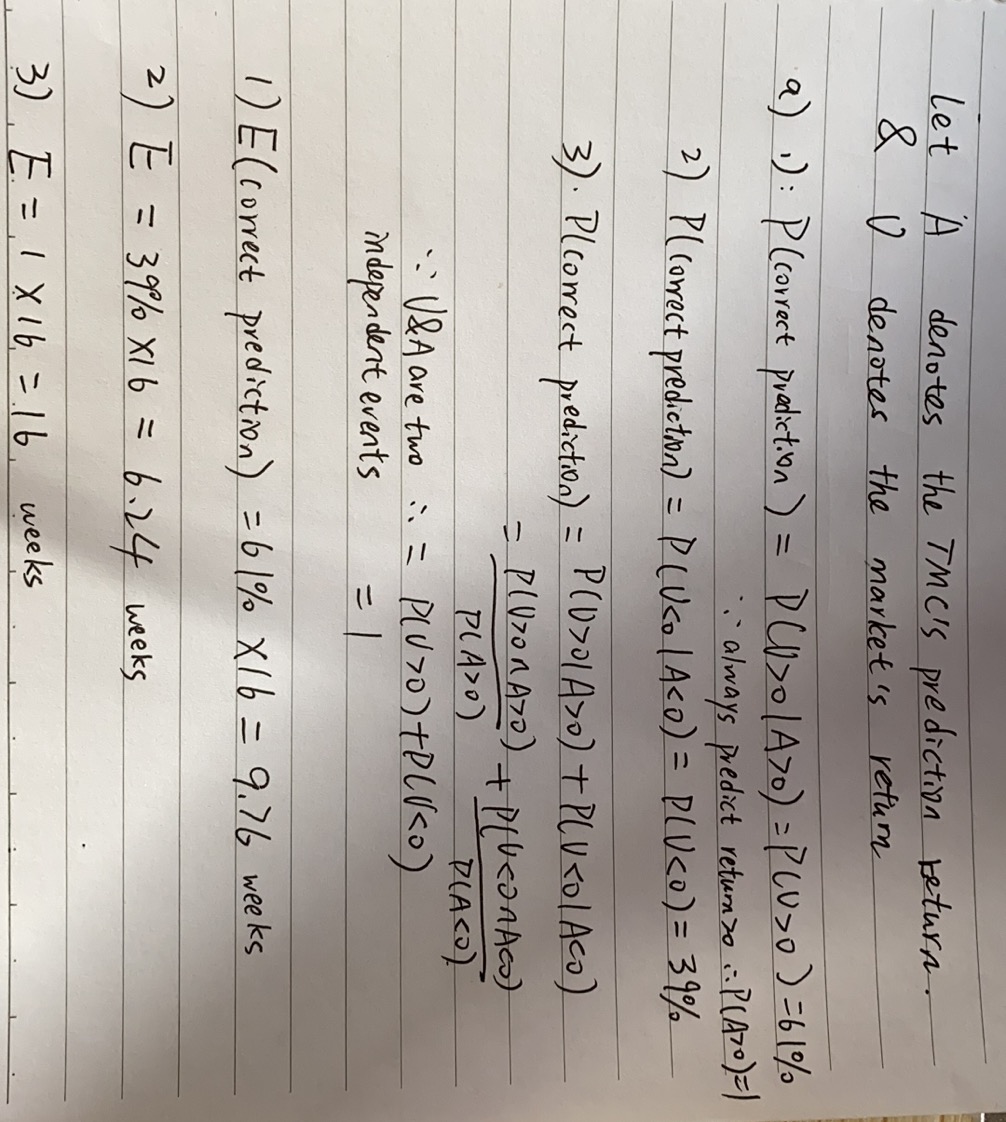
P = 1 – 3.6e-163

**Problem 2:** Conditional probabilities, total probability theorem, forecasting stock returns

The folks at TMZ Inc. (Two Mush Noize) don’t know much finance. They do know that people really care whether the market return is positive. They want to start a weekly financial letter which (seems to) predict the direction of the weekly S&P500 return. They consider three strategies: 1) always predict >0 return, 2) always predict <0 return, 3) randomly predicts >0 return 63% of the time.

Your published research shows that the weekly market return is positive with probability p(U)=61%, and is unpredictable. They must have found this interesting, so they hired you as a consultant for this project.

a) Give the weekly probability of success and the expected number of correct predictions after 16 weeks for each rule..



b) Explain in words, what is wrong with strategy 3) aka the “randomizing strategy”.

The events of randomizing strategy and the market return might have different distributions. The market return can be influenced by many factors like macro politics issues, fundamentals an so on..

c) TMZ is getting serious; they want to do conditional prediction. The rules in question a) were *unconditional* rules, you now investigate conditional rules, maybe markets are not efficient ! You will be able to charge them for some serious data analysis. The second column in the file “FF\_Research\_Data\_Factors\_weekly.csv” contains the S&P500 excess return over the risk free rate.[[1]](#footnote-1) Use only data from Jan. 2011 to Dec. 2019. Assume that future returns will behave consistently with these data.

Fill three versions of the two-way table: 1) simple counts 2) joint probabilities 3) conditional probabilities.

Table 1 Count Table

|  |  |  |
| --- | --- | --- |
|  | Rt < 0 | Rt > 0 |
| Rt-1 < 0 | 71 | 113 |
| Rt-1 > 0 | 164 | 168 |

Table 2 Joint Probability

|  |  |  |
| --- | --- | --- |
|  | Rt < 0 | Rt > 0 |
| Rt-1 < 0 | 15.17% | 24.15% |
| Rt-1 > 0 | 24.36% | 35.90% |

Table 3 Conditional Probability (Rt | Rt-1 )

|  |  |  |
| --- | --- | --- |
|  | Rt < 0 | Rt > 0 |
| Rt-1 < 0 | 38.38% | 61.08% |
| Rt-1 > 0 | 40.43% | 59.57% |

d) Use the numbers in the conditional table to give the **unconditional probability** of success of the rule that predicts that every week the S&P is up (down), it will go up (down).

The unconditional probability of success if predicts every week’s S&P is up is the sum of the second column in conditional table, aka 61.08% + 59.57% = 120.65%.

The unconditional probability of success if predicts every week’s S&P is down is the sum of the first column in conditional table, aka 38.38% + 40.43% = 78.81%.

e) Given your tables, can a conditional rule improve on this rule?

Yes, just change . This will cover all the situations might occur. Actually in the previous time domain, there is one data of return = 0 at 20140516. So adding return = 0 condition will include all the situations might occur and ensure the sum of joint probability table equals 1 and the sum of each row of the conditional probability table equals 1.

f) Given these results, how efficient do you think the US market is?

Not much efficient because one can get the probability of success more than one as long as he/she always predicts weekly S&P is up. (Efficient market means that one cannot beat the market because all the potential information has already been shared.)

g) Would you work for these people, why or why not ?

It depends on my research time domain and the research outcome. If the above is what my final outcome is, then I will not work for these people because after they see my research, they are very likely turn to their strategy 1 and hope for making easy money. But in reality, this may cause great loss and they might come to blame me afterwards. That’s bad.

**Problem 2:** Bayes rule and conditioning properly

The Tears Nobucks company is experimenting persistent decline in sales and profits. The North East region manager, Mrs Chopheads, has been instructed to shut down two of the three stores left in Metro Boston. She communicated to the store managers of Burlington (Mr Bean), Natick (Mr Natty), Quincy (Mrs Quince), that two of them will be fired and their stores closed, only one will keep his/her job. They all have the same probability of being fired: p(B)=p(N)=p(Q)=2/3.

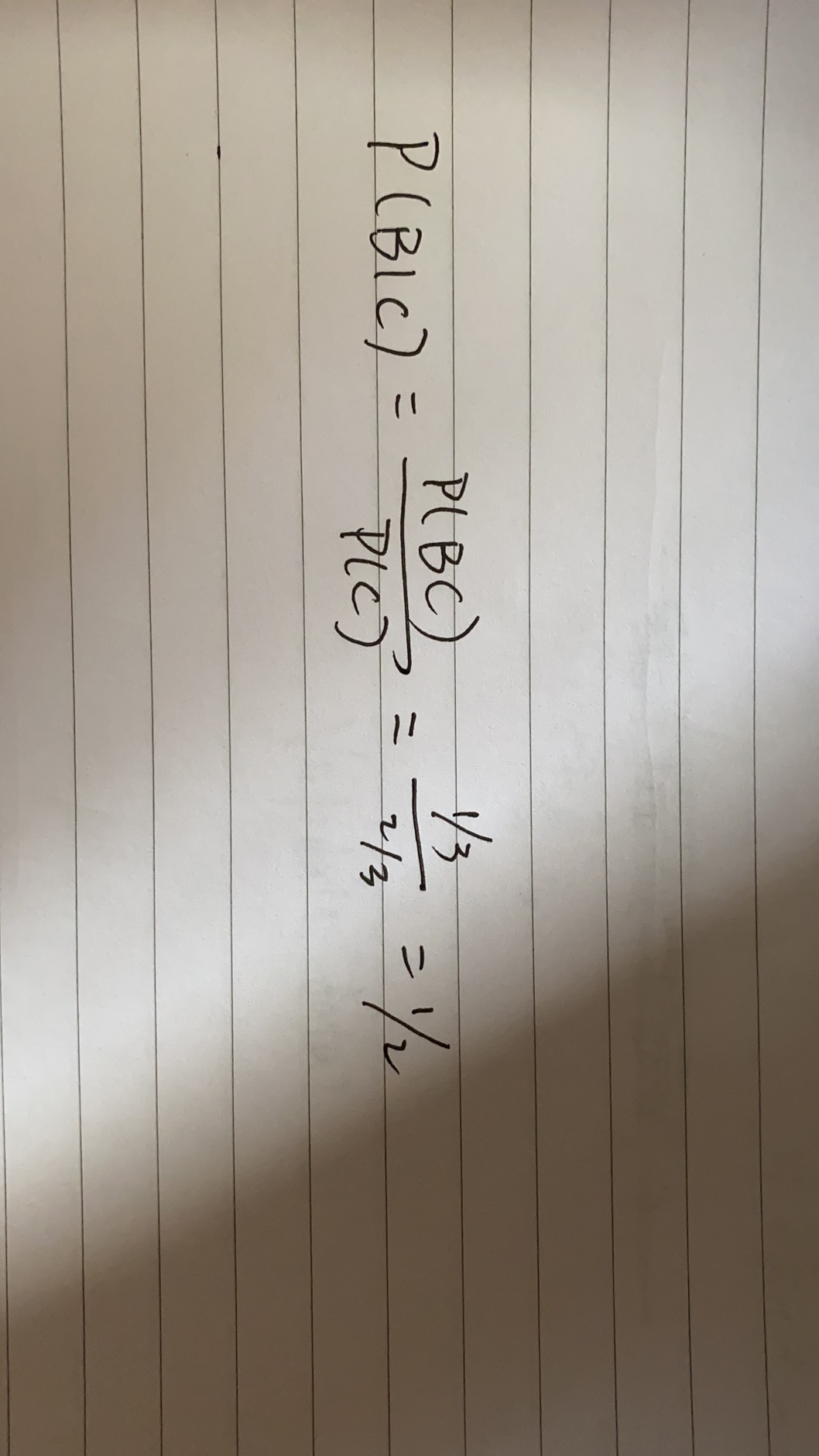
Mrs. Chopheads knows which store will remain open but obviously must not tell. At a virtual not-so happy hour (and maybe over one too many Summer Ales), she confided to the Burlington manager that the Quincy store would close.

* Mr Bean tells himself: Good news! I now know that Quincy will close, so either my store or Natick will close. My probability of being fired is only ½, not 2/3.
* Mrs. Chopheads realizes that she broke the rule of silence, but she tells herself: “Either Natick or Quincy must close anyway, so I have given Mr Bean no information on **his** store, so he should still think his probability of being fired is 2/3”.

Who is wrong, Mrs Chopheads or Mr Bean? It is a question of proper conditioning!

Call “C” the event: Ms Chopheads tells Mr Bean that the Quincy store will close.

a) Compute p(B|C). Was Mrs Chopheads or Mr Bean correct?



Mr Bean is correct. (We assume that Mrs. Chopheads is not lying here.)

b) What “wrong” conditioning “?” did Mr Bean use to come up with p(B| ?) = 1/2

Actually the event that store Q will be closed is not the same as the event C tell B that Q will be closed. There is potential lying possibility, which makes Mr. Bean’s conditioning not so proper.

**Problem 4:** Conditional probabilities, fund performance

Your boss wants a quick and not dirty recent performance analysis of mutual funds. She reads that most funds don’t beat the market but can’t get a recent review of fund performance.

You collected monthly returns from Jan. 2010 to Dec. 2017 for 1500 funds fully invested in the stock market, file 1500fund-monret.csv in the Data folder. It contains monthly returns on 1500 mutual funds from 1/2010 to 12/2017.

You know you can go to Ken French’s data web site and get the monthly return on the market index for the exact same period, so you do it!

a) Compute the **average monthly return** for each fund for the period (2010/6-2013/12), and the market.

* In Fig. 1 Plot a histogram of these 1500 average fund returns. Annualize these monthly averages by multiplying them by 12 so they have kind of an annual magnitude to them. Add a vertical bar in black for the average **of the 1500 averages**, and a vertical bar in blue for the market average.

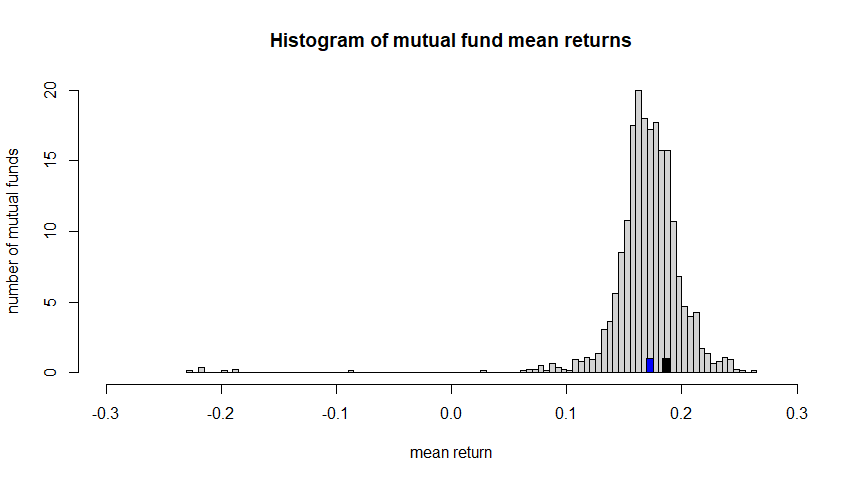


Fig.1 Histogram of mutual fund mean returns

* What % of funds beat the market for that period?

29.2 %

* what would you expect the result to be if the **fund managers were** randomly picking stocks ?

During this period, there are expected 29.2% stocks that could beat the market return. ( regardless of the market cap)

* In a couple **sentences** tell your boss what you think of mutual fund performance for the first period.

In the first period, only 29.2% of mutual funds can beat the market return on average. So the overall performances are not good at all.

b) Your boss says that this is all dandy but she wants you to find out if even a small number of funds can beat the market consistently. “Ahah” you say, “this is why I saved the 2014-2017 period, I will prepare a persistence analysis using basic concepts of joint and conditional probabilities”. Are some funds consistently the best?

* As a first pass at the problem, compute Period 1 (2010/6-2013/12) and Period 2 (2014-17) mean returns for each fund. Plot R2 vs R1. This scatter plot has 1500 points.

Add the Market as a point to the graph; choose a symbol that makes it visible (see help(“par”) and help(“points”) in R)

Add a vertical and a horizontal dashed lines going through the market (see abline and lty graphic parameter in R)

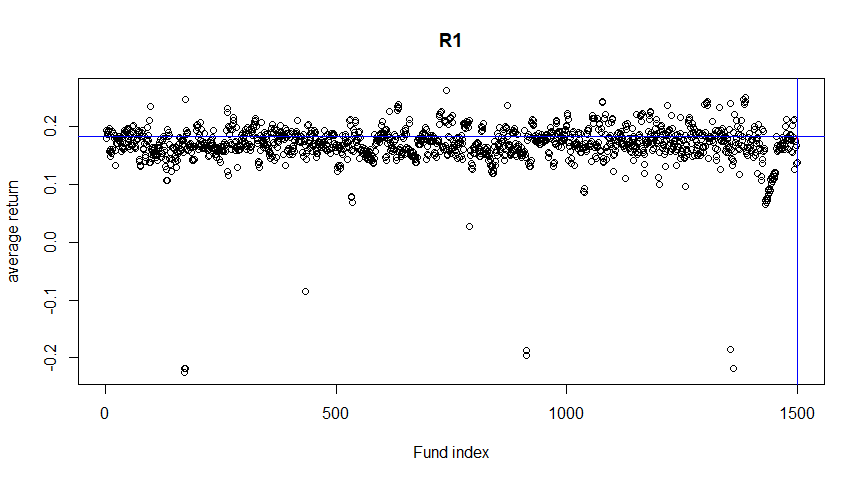


Fig 2. Scatter plot of mutual funds’ returns in period 1

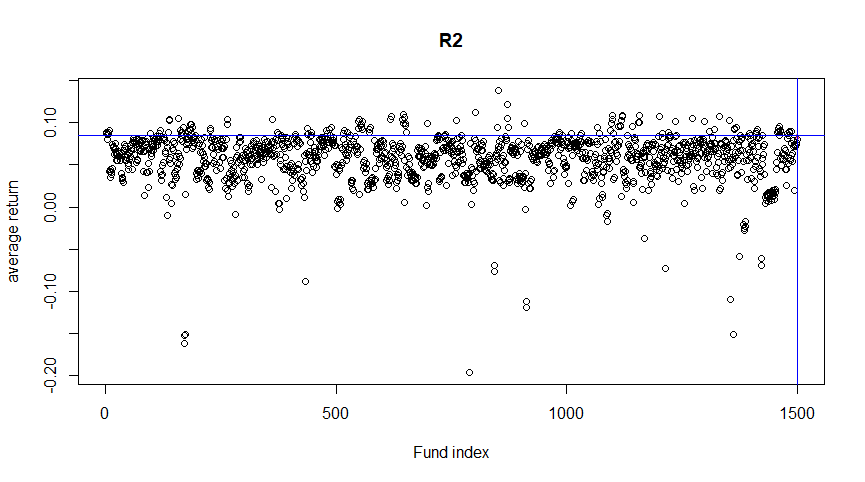


Fig 3. Scatter plot of mutual funds’ returns in period 2

Do you see any pattern in this plot?

Yeah, most data points concentrate in some intervals.

What would the plot look like if there was no persistence in performance?

No obvious concentration of the data points.

* Define W (win) as a fund being top 20%, L (lose) bottom 20%, and M (middle) the middle 60% range. Fill three versions of the two-way table for period 2 vs 1: 1A) simple counts out of 1500 1B) joint probabilities 1C) conditional probabilities. Keep only 2 digits for probabilities

Table 1A: Persistence in fund ranking, total counts

|  |  |  |  |
| --- | --- | --- | --- |
|  | L2 | M2 | W2 |
| L1 | 136 | 142 | 22 |
| M1 | 89 | 608 | 203 |
| W1 | 75 | 150 | 75 |

.

Table 1B: Persistence in fund ranking, joint probabilities

|  |  |  |  |
| --- | --- | --- | --- |
|  | L2 | M2 | W2 |
| L1 | 9.07% | 9.47% | 1.47% |
| M1 | 5.93% | 40.53% | 13.53% |
| W1 | 5.00% | 10.00% | 5.00% |

.

Table 1C: Persistence in fund ranking, conditional probabilities 2|1

|  |  |  |  |
| --- | --- | --- | --- |
|  | L2 | M2 | W2 |
| L1 | 45.33% | 47.33% | 7.33% |
| M1 | 9.89% | 67.56% | 22.56% |
| W1 | 25.00% | 50.00% | 25.00% |

.

\*\*\* Here we assume the break point belongs to the lower interval\*\*\*

(e.g. If the fund locates just at the 20% top point, then it is assumed in middle.

and if the fund locates just at the 20% bottom point, it is assumed in lose.)

* Add Table 1D, to show what Table 1C should be if funds had no different abilities or persistence in abilities.
* In a few sentences summarize these results to your boss, explaining the evidence on the ability of the best funds to remain the best, the so-called “hot-hand”.

c) This was a nice effort but your boss is still not impressed. “Anyway, what do I care that the best remain the best if they can’t reliably beat the market! I will still advise our clients to buy DFA or Vanguard! Do me a table showing persistence in beating the market”

* This time you prepare a two-by two set of three tables 2A, 2B, 2C,2D. You define Win / Lose by having a larger/ smaller return than the market.

Table 2A: Persistence in Beating the market Counts of Funds

|  |  |  |
| --- | --- | --- |
|  | L2 | W2 |
| L1 | 979 | 83 |
| W1 | 386 | 52 |

Table 2B: Persistence in Beating the market joint probabilities

|  |  |  |
| --- | --- | --- |
|  | L2 | W2 |
| L1 | 65.27% | 5.53% |
| W1 | 25.73% | 3.47% |

Table 2C: Persistence in Beating the market conditional probabilities 2|1

|  |  |  |
| --- | --- | --- |
|  | L2 | W2 |
| L1 | 92.18% | 7.82% |
| W1 | 88.13% | 11.87% |

1A) simple counts out of 1500 1B) joint probabilities 1C) conditional probabilities.

* Conclude with respect to persistence in ability to beat the market.

The persistence in ability to beat the market is just P(W2|W1) = 11.87%, which is quite low. So these mutual fund managers in general don’t have such persistence in ability to beat the market.

Do not answer this question but start working on it for homework 2

You total wealth is $10,000. A project can earn 30% or lose 10% with probability 0.5. Your utility of wealth has the shape U(W) = -1/W. You consider whether to invest your total wealth into it!

a) What is the $CE of your total wealth if you undertake the project? Do you do it?

b) You can borrow $B at 0%. What is the maximum Bmax you can borrow without going bankrupt in the down case. In Figure 3 plot your EXPECTED wealth vs the amount borrowed for B in [0 , Bmax]

c) Now write the simple formula of your $CE as a function of $B. In Figure 1, Use R to plot $CE vs the $B. Have B go from $0 to $Bmax.

d) How much would you borrow optimally to increase your expected utility? Show that point on your Figure 1.

Appendix (R code):

# problem 1.c

pro <- choose(99,3)\*choose(360,1)\*choose(359,96)/ (360\*\*99)

# problem 1.d

pro\_1 <- 1 - choose(360,99)/(360\*\*99)

# set directory & read data from file

setwd("C:/Users/jinji/Desktop/MF793File")

array <- read.csv("FF\_Research\_Data\_Factors\_weekly.csv")

return\_array <- array[array[,1] >= 20110101 & array[,1] <= 20191231,]

r\_i <- return\_array[2:length(return\_array[,1]), c(2)]

r\_i\_1 <- return\_array[1:length(return\_array[,1])-1, c(2)]

mat\_return <- cbind(r\_i\_1, r\_i)

# problem 2.c.1

a <- sum(mat\_return[,1]<0 & mat\_return[,2]<0)

b <- sum(mat\_return[,1]<0 & mat\_return[,2]>0)

c <- sum(mat\_return[,1]>0 & mat\_return[,2]<0)

d <- sum(mat\_return[,1]>0 & mat\_return[,2]>0)

# problem 2.c.2

# The sum of all four probabilities not equal 1 because exist one data = 0 at

# 20140516 in the time domain

sum <- length(mat\_return[,1])

a\_1 <- a/sum

b\_1 <- b/sum

c\_1 <- c/sum

d\_1 <- d/sum

# problem 2.c.3

# The sum of first row not equal 1 because exist one data r(i-1)<0 r(i)=0

a\_2 <- a/sum(mat\_return[,1]<0)

b\_2 <- b/sum(mat\_return[,1]<0)

c\_2 <- c/sum(mat\_return[,1]>0)

d\_2 <- d/sum(mat\_return[,1]>0)

# problem 4.a

fund\_array\_raw <- read.csv("1500fund-montrets.csv")

market\_array\_raw <- read.csv("FF\_Research\_Data\_Factors\_monthly.csv")

fund\_array <- fund\_array\_raw[fund\_array\_raw[,2]>20100531 & fund\_array\_raw[,2]<20140101,]

market\_array <- market\_array\_raw[market\_array\_raw[,1]>201005 & market\_array\_raw[,1]<201401,]/100

total\_array <- cbind(fund\_array, market\_array[,2])

mean\_array\_raw <- total\_array[,3:dim(total\_array)[2]]

mean\_array <- colMeans(mean\_array\_raw[sapply(mean\_array\_raw, is.numeric)])

fund\_mean\_array <- mean\_array[1:length(mean\_array)-1]\*12

hist(fund\_mean\_array, xlim = c(-0.3,0.3) , nclass = 100 , prob = T, xlab = "mean return", ylab = "number of mutual funds", main = "Histogram of mutual fund mean returns")

average\_market <- mean\_array[length(mean\_array)]\*12

average\_funds <- mean(fund\_mean\_array)

rect(average\_market, 0, average\_market + 0.6/100, 1, col = 'Black', add = TRUE)

rect(average\_funds, 0, average\_funds + 0.6/100, 1, col = 'Blue', add = TRUE)

percen\_beat\_market <- sum(fund\_mean\_array > average\_market)/length(fund\_mean\_array)

# problem 4.b

fund\_array\_2 <- fund\_array\_raw[fund\_array\_raw[,2]>20140531 & fund\_array\_raw[,2]<20170101,]

market\_array\_2 <- market\_array\_raw[market\_array\_raw[,1]>201405 & market\_array\_raw[,1]<201701,]/100

total\_array\_2 <- cbind(fund\_array\_2, market\_array\_2[,2])

mean\_array\_raw\_2 <- total\_array\_2[,3:dim(total\_array\_2)[2]]

mean\_array\_2 <- colMeans(mean\_array\_raw\_2[sapply(mean\_array\_raw\_2, is.numeric)])

fund\_mean\_array\_2 <- mean\_array\_2[1:length(mean\_array\_2)-1]\*12

average\_market\_2 <- mean\_array\_2[length(mean\_array\_2)]\*12

# plot in chart

x\_1 <- 1:length(fund\_mean\_array)

x\_2 <- 1:length(fund\_mean\_array\_2)

plot(x\_1, fund\_mean\_array, xlab = "Fund index", ylab = "average return", title("R1"))

points(length(x\_1)+1, average\_market, col ='red', pch = '\*')

abline(h=average\_market, col = 'Blue')

abline(v=length(x\_1)+1, col = 'Blue')

plot(x\_2, fund\_mean\_array\_2, xlab = "Fund index", ylab = "average return", title("R2"))

points(length(x\_2)+1, average\_market\_2, col ='red', pch = '\*')

abline(h=average\_market\_2, col = 'Blue')

abline(v=length(x\_2)+1, col = 'Blue')

sort\_fund\_mean\_array <- sort(fund\_mean\_array)

L\_1 <- sort\_fund\_mean\_array[0.2 \* length(sort\_fund\_mean\_array)]

M\_1 <- sort\_fund\_mean\_array[0.8 \* length(sort\_fund\_mean\_array)]

sort\_fund\_mean\_array\_2 <- sort(fund\_mean\_array\_2)

L\_2 <- sort\_fund\_mean\_array\_2[0.2 \* length(sort\_fund\_mean\_array\_2)]

M\_2 <- sort\_fund\_mean\_array\_2[0.8 \* length(sort\_fund\_mean\_array\_2)]

fund\_mean\_both <- rbind(fund\_mean\_array, fund\_mean\_array\_2)

# row 1

sum(fund\_mean\_both[1,]<=L\_1 & fund\_mean\_both[2,]<=L\_2)

sum(fund\_mean\_both[1,]<=L\_1 & fund\_mean\_both[2,]>L\_2 & fund\_mean\_both[2,]<=M\_2)

sum(fund\_mean\_both[1,]<=L\_1 & fund\_mean\_both[2,]>M\_2)

# row 2

sum(fund\_mean\_both[1,]>L\_1 & fund\_mean\_both[1,]<=M\_1 & fund\_mean\_both[2,]<=L\_2)

sum(fund\_mean\_both[1,]>L\_1 & fund\_mean\_both[1,]<=M\_1 & fund\_mean\_both[2,]>L\_2 & fund\_mean\_both[2,]<=M\_2)

sum(fund\_mean\_both[1,]>L\_1 & fund\_mean\_both[1,]<=M\_1 & fund\_mean\_both[2,]>M\_2)

# row 3

sum(fund\_mean\_both[1,]>M\_1 & fund\_mean\_both[2,]<=L\_2)

sum(fund\_mean\_both[1,]>M\_1 & fund\_mean\_both[2,]>L\_2 & fund\_mean\_both[2,]<=M\_2)

sum(fund\_mean\_both[1,]>M\_1 & fund\_mean\_both[2,]>M\_2)

# problem 4.c

fund\_mean\_mkt\_premium <- fund\_mean\_array - average\_market

fund\_mean\_mkt\_premium\_2 <- fund\_mean\_array\_2 - average\_market\_2

fund\_mean\_mkt\_premium\_both <- rbind(fund\_mean\_mkt\_premium, fund\_mean\_mkt\_premium\_2)

# first row

sum(fund\_mean\_mkt\_premium\_both[1,]<0 & fund\_mean\_mkt\_premium\_both[2,] <0)

sum(fund\_mean\_mkt\_premium\_both[1,]<0 & fund\_mean\_mkt\_premium\_both[2,] >0)

# second row

sum(fund\_mean\_mkt\_premium\_both[1,]>0 & fund\_mean\_mkt\_premium\_both[2,] <0)

sum(fund\_mean\_mkt\_premium\_both[1,]>0 & fund\_mean\_mkt\_premium\_both[2,] >0)

1. It is a good time to go on Ken French’s web page and look at all the data available there. You can’t live in the data based portfolio management space without knowing this data source. There you will find explanations on the data and collection process. Always read the details and description of the data [↑](#footnote-ref-1)