Boston University Questrom School of Business
MF 793 – Fall 2017

Eric Jacquier

Problem Set 1 - Solution
Due Monday September 24th in class

Problems turned in after the beginning of student section have a notch deduction.
Problems turned in after class get a zero.

- Do the Problem Set in groups of two – both students in the same section
- Turn in one paper copy in class with two names, no electronic submission accepted.
- **To get a check, you need to answer all the questions.**
- **Discussion questions must be hand written with a pen to count**.
- R help at the bottom of this document
- All Figures must be professionally made with X and Y axes labels and numbers. Tables must have row and column names and a Title. The number 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.

Problem 1: Counting

The new middle school class has arrived, all 230 of them. The social events director wants to know the probability that exactly 4 kids (no more, no less) have their birthday the same day, and the rest ALL have different birthdays. Use a 365 day year.

a) Write by hand the theoretical formula. Explain in words how you came up with each piece (the numerator, the denominator)

b) Compute it in R and give the result. Show the code you used.

c) Also what would be the probability that at least (possibly more) 2 students share a birthday? Give the formula and the final result

a) $\dfrac{\binom{230}{4} * 365 * 364 * \ldots * (365-226)}{365^{230}}$

Numerator: what 4 out of the 230 have the same birthday, times what day it is. Then the remaining 226 students can have their separate birthday on 364 days for the first, 363 days for the second, etc…
Denominator: Total combinations possible. Each student can have its birthday (or not) on any of 365 days.

b) In the numerator the factorial of 365 overflows. The prod(139:365) function also overflows. The denominator also overflows.
We can use the log of the expression and then exponentiate it. The command lfactorial in R directly computes the log of a factorial, avoiding the overflow. We can also log everything else

The way to compute a ratio of overflow numbers is to go through the log of each part and then exponentiate.

We often need to compute density ordinates which include the gamma function.
Recall (you need to know this result) that: **Γ (n) = Factorial(n-1)**
.. even though the gamma is a function on the real line and factorial is for integers.
In that setup, n is often a sample size. R has the log of the gamma function: **lgamma(x)**

Result: 1.395459e-42

c) This is the easiest birthday question! It is one minus the probability that none of them share a birthday:

1- [365 * ... * (365-229+1)] / 365^{230} = 1
There are about no chances at least 2 students share a birthday

Problem 2: Bayes rule and conditioning properly

The Tears Nobucks company is experimenting persistent decline in sales and profits. The NorthEast 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 not-so happy hour (and maybe over one too many Summer Ales), she told in confidence 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 B 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 B correct?
b) What "wrong" conditioning "?" did Mr B use to come up with p(B| ?) = 1/2

a) Bayes: P(B|C) = P(C|B) P(B) / P(C) So we need to compute P(C)

Total Probability Theorem for the denominator as usual but **potential mistake**:
Can we write: P(C) = p(C|B) p(B)+ p(C|Q) p(Q)+ p(C|N) p(N) ?
NO! Because B, N, and Q are overlapping events since 2 stores will be closed.

To use the Total probability theorem we must use “keeping the job” as an event. These are non-overlapping since only one can keep her job.
So, “not B” means Mr Bean is not fired

P(C) = $p(C \mid \text{not } B)\ (1-p(B)) + p(C \mid \text{not } Q)\ (1-p(Q) + p(C \mid \text{not } N)\ (1-p(N))$
$= [\ 1/2 \quad + \quad 0 \quad + 1 \quad] * 1/3$ = **1/2**.

$p(C \mid \text{not } N) = 1$ because, if N does not close, Mrs Chopheads tells Mr B that Q will be fired
$p(C \mid \text{not } Q) = 0$ clearly !
Now we can finish Bayes Theorem:

P(B|C) = $p(C|B)\ p(B) / p(C) = (1/2)*(2/3) / (1/2)$ = **2/3**

Mrs Chopheads was correct, she revealed nothing to Mr Bean,
his probability of being fired is still 2/3.

b) $p(B|Q) = p(Q|B)p(B)/p(Q) = (1/2) * (2/3) / (2/3) = 1/2$!
Whether Mr Bean knows Bayes Theorem or not, a simple error he most likely made was to confuse two different events
“ Quincy will close”
“Mrs Chopheads tells him that Quincy will close”

Problem 3: Was it a mule or a horse?

A town has two taxi companies. Blue Mules runs 25 dark blue cabs. Dark Horse operates 75 grey cabs. One dark foggy night, a taxi is involved in a hit-and-run accident. The town's taxis of both companies were all on the streets at the time of the accident. A witness claims to have seen a blue taxi drive away from the scene.
The inspector in charge, Mr Cluesoh remembers Bayes theorem. He makes the witness take a vision test in dark conditions. Presented repeatedly with random blue or grey cars, she correctly identifies grey cars 17 out of 20 times and blue cars 7 out of 10 times.

Mr Cluesoh has now all what’s needed to compute the odds ratio that a Blue Mule did it vs a Dark Horse did it. He turns the data over to you to finish the job.

Use the notation for the events: M: a Blue Mules taxi did it, H: a Dark Horse taxi did it, SM: Witness sees a blue taxi. SH: Witness sees a grey taxi.

Compute p(M|SM) and p(H|SM), make sure to show all intermediate steps.

Look up the definition of an odds ratio and compute the odds ratio Inspector Cluesoh needs.

This is a straightforward use of the theorem.
We want $p(M \mid SB)$! Let’s do the denominator first by total probability theorem

P(SM) = $P(SM \mid M) \quad P(M) + P(SM|H) \quad P(H)$ clearly $p(M) = 1 - p(H)$

= (7/10) (25/100) + (3/20) (75/100)
= 0.2875

P(M | SM) = P(SM|M) P(M) / P(SM) = 0.7 * 0.25 / 0.2875
= **0.61** Yikes, only 60% chances !!!

P(H | SM) = 0.39

An **odds ratio**, as fans of horse racing know, is the simple **ratio of the two probabilities**. Here we have

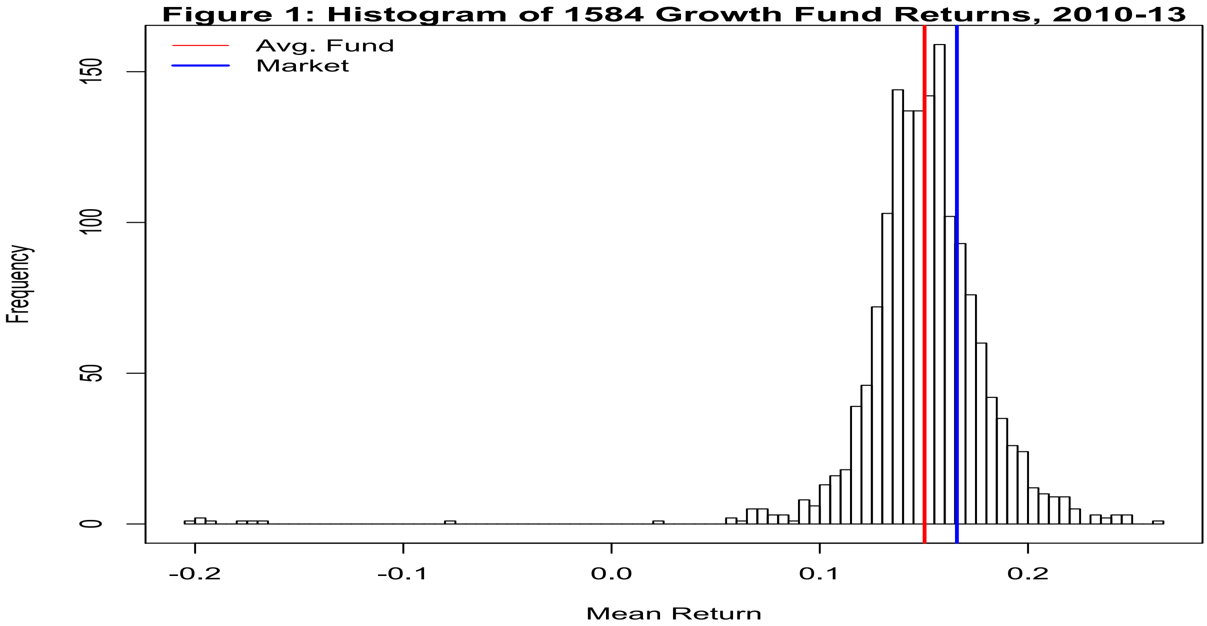
ODD (M / H) = 1.55 **This is not considered particularly strong evidence**

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.
Lucky enough, you still have your MF793 file 1584gfund-monret.csv in the Resources/Data folder. It contains monthly returns on 1584 mutual funds from 1/2010 to 12/2017 . The funds are largely institutional, domestic, growth funds.
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 4 year ***average monthly return*** *for each fund (2010-13), and the market.*

- *In Fig. 1 Plot a histogram of these 1584 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 1584 averages****, and a vertical bar in red for the market average.*
- *What % of funds beat the market for that period?*
- *If the* ***fund managers were*** *randomly choosing stocks what would you expect the result to be?*
- *In a couple* ***sentences*** *tell your boss what you think of mutual fund performance for the period.*



- 393 out of 1584 beat the market over 2010-2013. That is 25% of the funds

- If fund managers had no special ability, for example were picking stocks randomly, we would expect 50% of them to beat the market.

- The message to the boss is that this group of 1584 growth funds perform worse as a group than one would expect if they were just randomly picking stocks. It might be that by selling and buying stocks often they incur transaction costs that further reduce their performance relative to the market. All in all, a pretty bad result.

- A very small number of funds had a horrendous performance. How can one have < 0 average returns while at the same time the market return was: 18%, 1.7%, 16%, and 30% in 2010-11-12-13? If is so bad it calls for checking for data errors.

b) Your boss tells you that this is all fine but maybe a small number of funds can beat the market consistently. "Ahah" you say, "this is why you saved the 2014-2017 period, you 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 0 (2010-13) and Period 1 (2014-17) mean returns for each fund. Plot R1 vs R0. This scatter plot should have 1584 points on it. Add the Market as a point to the graph, make sure to choose a symbol that makes it visible (see help("par") and help("points") in R)*
  Do you see any pattern in this plot?
  What would the plot look like if there was no persistence in performance?

The boss is right. If these 25% are persistent performers, that is if we can predict from period to period that the top funds are more likely than not to remain the top funds, this is still interesting evidence.

With no persistence at all, the plot should look like a circular scatter with no evidence that high means in P1 correspond to high means in P2. There seems to be some mild evidence of persistence since the plotted regression line is not horizontal.

Note how the poor performers increase the impression of persistence. Removing them reduces the strength of evidence.

We know that means are measured with a **LOT** of sampling error. So we can do a "robust" version of this plot by just looking at the ranks of the funds.

Figure 2: Fund mean returns, second vs first period returns

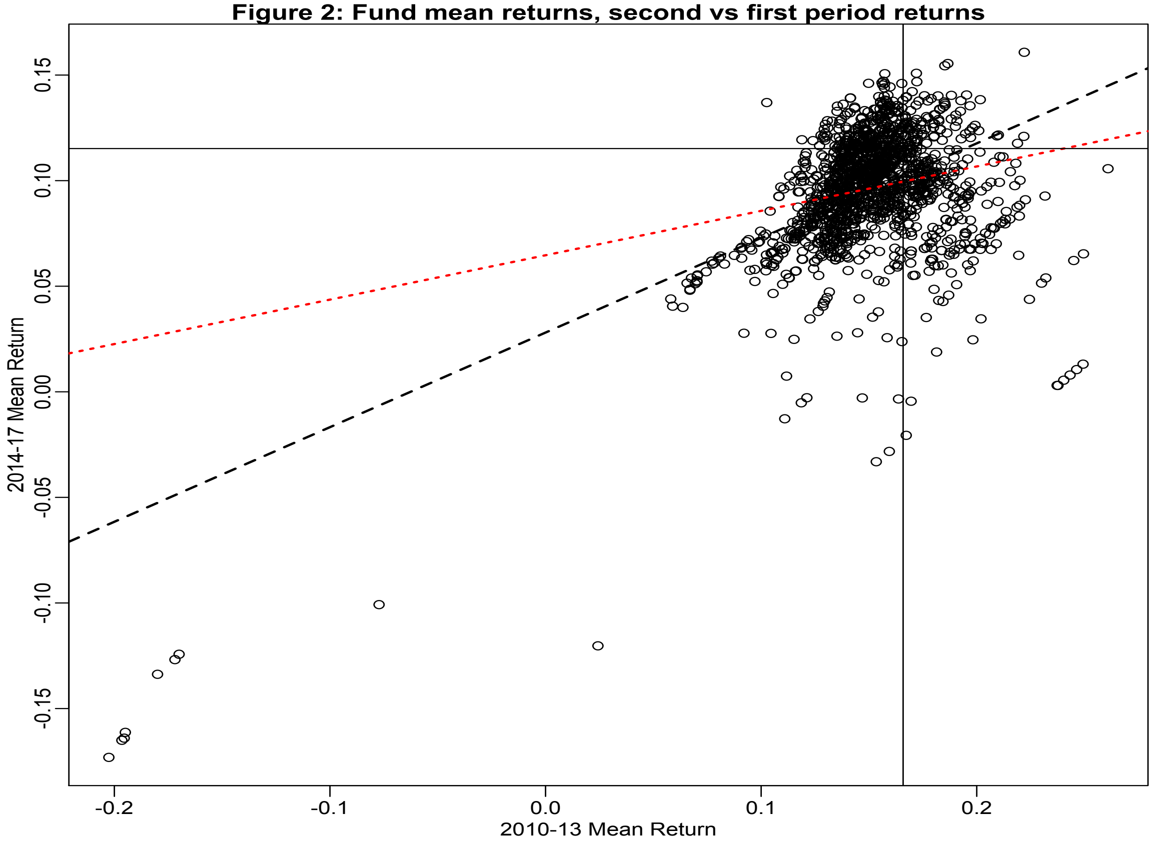
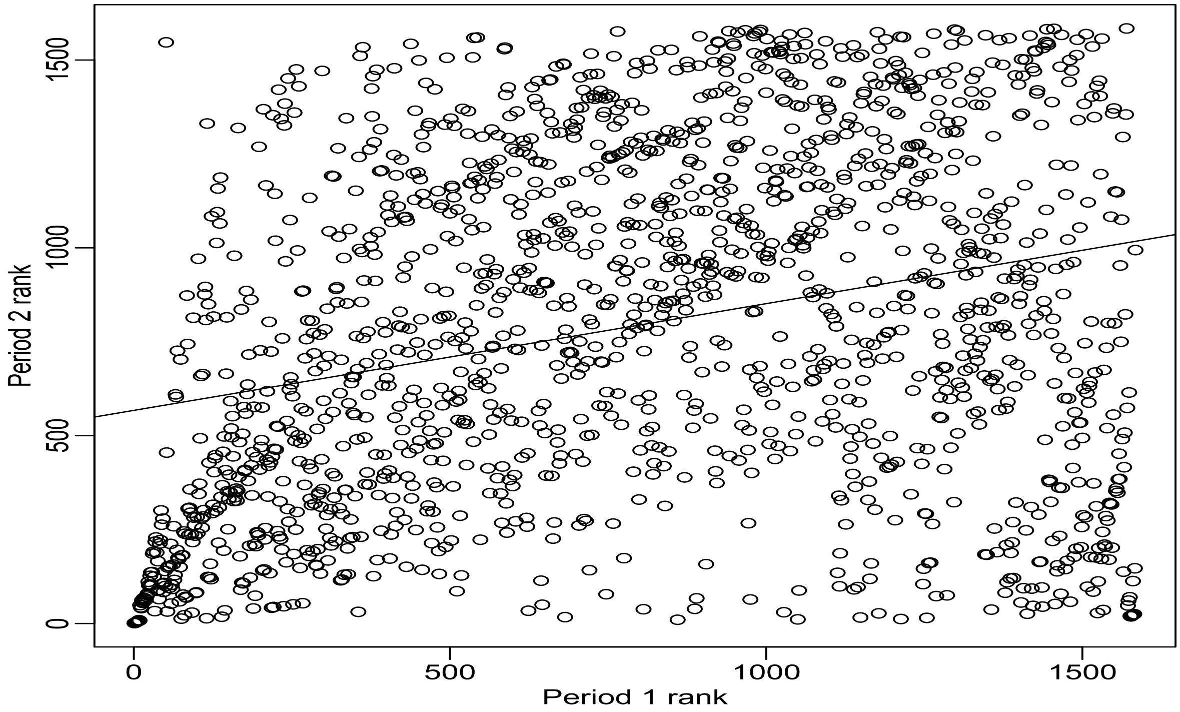


Figure 2b: Funds rank persistence from 10-13 to 14-17



- *Define W (win) as a fund being top 20%, L (lose) bottom 20%, and M (middle) the middle 60% range. Period 0 is 2010-13, period 1 is 2014-17. Fill three versions of the two-way table. That will be Tables 1A, 1B, 1C. Keep no more than 2 digits for probabilities*

Table 1A: Persistence in fund ranking, counts

	L1	M1	W1
L0	**158**	144	15
M0	87	632	231
W0	72	174	71

Table 1B: Persistence in fund ranking, joint probabilities

	L1	M1	W1
L0	**0.10**	0.09	0.01
M0	0.05	0.40	0.15
W0	0.05	0.11	0.04

Table 1C: Persistence in fund ranking, conditional probabilities

	L1	M1	W1
L0	**0.50**	0.45	**0.05**
M0	0.09	0.66	**0.24**
W0	0.23	0.55	**0.22**

Table 1D: Conditional probabilities with no persistence

	L1	M1	W1
L0	0.2	0.6	0.2
M0	0.2	0.6	0.2
W0	0.2	0.6	0.2

- *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".*

The evidence is thin at best! The conditional probability of being top 20% in period 2 given one was in period 1 is 0.22, about equal to the probability conditional on being in the middle in period 1that is undistinguishable from the unconditional probability. Loses however seems persistent ! Period 1 bottom funds have a 50% chance of being bottom again, and only 5% chance of being top. It's a shame we can't short sell mutual funds to exploit the losers' hot hand.

Now we know the regression lines on the scatter plots have positive (but weak) slopes. They are driven by the persistence in losers, not by any persistence in winners.

c) This was a nice effort but your boss is not impressed. "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"

Note that given the results above, we don't expect much from this additional analysis. However let's do it. It's simpler with only one cutoff.

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

We already found that 25% of the funds (393) beat the market in 2010-13.
Only 22% (344) beat the market in 2014-17.

Table 2A: Persistence in Beating the Market, Counts

	L1	W1
L0	945	246
W0	295	98

Table 2B: Persistence in Beating the Market, Joint Probabilities

	L1	W1
L0	0.60	0.15
W0	0.19	**0.06**

Table 2C: Persistence in Beating the Market, Conditional Probabilities

	L1	W1
L0	0.79	0.21
W0	0.75	**0.25**

Table 2D: Conditional Probs, no persistence

	L1	W1
L0	0.75	0.25
W0	0.75	0.25

Table 2E: Joint Probs, No Persistence

	L1	W1
L0	0.56	0.19
W0	0.19	0.06

Two possible premises for Tables 2DE. You could use 50% as the unconditional 1-period probability, then each cell would be 50%. One could also use no persistence based on a 25% unconditional probability of beating the market as observed in period 1. We used that second interpretation.

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

In b) we defined performance as being the top. In c) we define performance as beating the market. Again, we find **NO** persistence in performance for the top funds. Their conditional probability of beating the market is equal to the unconditional probability. Let's say it another way. If you pick a fund that beat the market this past period – 25% of them did, you have 25 % chances it beats the market next period !

d) Your boss is interested. Then she says: " Maybe these are not the best results possible. Why did you decide on top 20% as winning cut off, why not top 10% or top 30%. Can't you find the cutoff that will show the best (persistent) performance? Same for the market, why don't you look at the funds that beat the market by some margin to search for persistence. Maybe the better funds with persistence beat the market with some margin regularly".

You remember reading a very interesting article on this topic the first week of class in MF793. Write a convincing couple paragraphs answering your boss using concepts from that article. Make explicit quotes from the article. It must be written by hand and legible to count for your report.

The main concept that applies here is "data mining". The term "p-hacking" in the article refers to the practice of searching for the most convincing set of explanatory variable and then reporting the p-value (another way to express the magnitude of the t-statistic estimated). The problem is the the p-value is not designed to take data mining into account.

Try to explain it this way to your boss: Think of trying 100 explanatory variables (X_1 ….. X_{100}) to predict a dependent variable Y. Under the null hypothesis that none of these explain Y, we still expect to randomly find 5% of them that have a t-statistic rejecting the null at the 5% level. In fact that is exactly the definition of the "size of the test": If we test at the α% level, we will incorrectly reject the null 5% of the time.
Things go wrong when we data-mine and don't take it into account. For example, we find 5 X_i variables and write a paper or a memo on how important these variables are without telling the reader that we earched from 100 candidate variables (data mining, aka data snooping). Finding 5 variables with significant t-stats out of a 100 is exactly what we would expect under the null hypothesis that none of these variables is a predictor for Y.

There are statistical techniques for adjusting for data mining (such as Bonferoni adjustments, out-of-sample checks, etc…). But the vast majority of people who conduct data mining do not use them.

So, how does it fit our case. The data mining here would be over the cut-off x (Top x% of the funds) and report only for the cut-off which has the strongest evidence of persistence. Given the randomness of data, for any sample, say we search from top 50% to top 1% by increment of 1%, there will be a cutoff with the best results. And this cutoff may appear to reject the null hypothesis of no persistence.

One approach would be to simulate the data mining under the null: Simulate lots of samples of fund returns with no persistence ability. For each sample search the cut off that shows the most persistence. This would allow us to have the distribution of the data-mining driven evidence of persistence under the null of no-persistence! We could then use that distribution to conclude whether we reject the null. That would be correct. Ummh, this would be a nice project idea.

***Problem 5**: Attitude to risk and return*

Your 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 B_{max} 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 , B_{max}]
c) Now write the simple formula of your \$CE as a function of \$B. In Figure 4, Use R to plot \$CE vs the \$B. Have B go from \$0 to \$$B_{max}$.
d) How much would you borrow optimally to increase your expected utility? Show that point on your Figure 4.

a) $CE = U^{-1}(EU(W)) = -1 / [\ 0.5 * (-1/13000) + 0.5 * (-1 / 9000)\ ]$

$= \$\ \mathbf{10636}$

So we undertake the project which will raise our CE by \$636

b) Obviously easy to generalize this problem to a positive interest rate.

Net Worth in the down case: $(B+10{,}000)*0.9 - B = 9{,}000 - 0.1\,B \geq 0$

Xan't borrow more than: $\mathbf{B_{MAX} = \$\ 90{,}000}$

$EW = 0.5\ [(B+10{,}000) * 1.3 + (B+10{,}000) * 0.9\ ] - B$

$\mathbf{EW = \$\ 11{,}000 + 0.1\ B}$ Expected wealth is clearly unbounded as we borrow more.

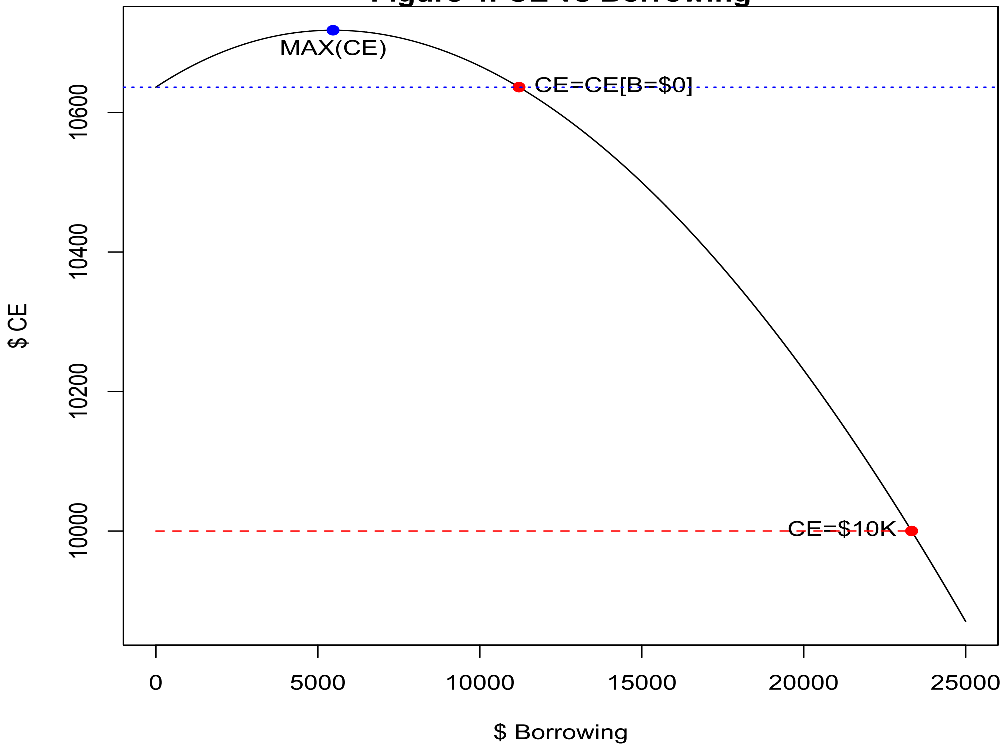
Only on account of EW, we would borrow as much as we are allowed to.

c) Because of risk aversion, we will not do this.

$CE(W) = -1 / [\ 0.5 * (-1/(13000+0.3B) + 0.5 * (-1 / (9000-0.1B)\ ]$

See Rcode for basic plotting tricks

Figure 4: CE vs Borrowing



The Following R commands are of interest

- Do help(“commandname”) for details on how to use a command

- Counting: choose, factorial, lfactorial

  ```
  # What if you need factorial(171) or factorial(172) .... or bigger?
  # For ratios of overflow numbers, make partial divisions and take the product after
  # or remember that: ratio = exp(log(ratio)), and log(ratio) = log(numerator)-log(denominator)
  ```

- Reading Data

  The easiest is simple rectangular data sets with a header row in .csv format

  ```
  arrayname <- read.csv(“filename.csv”)
                          # assumes a header, first row contains name variables
  also: read.table(), scan()
  ```

- Writing Data

  Think of putting table results in a R array, then write the array to a .csv file. The .csv file can then be put directly in a .xls table and into a word document
  There are many ways to prepare latex tables in R for those using that more sophisticated quantitative paper formatting package.

  ```
  write.csv(arrayname, “table.csv”)

  also: write.table().  write()
  ```

- Choosing subsets of data

  Say we have a matrix with 200 rows and 5 columns. We can select any subset we want

  ```
  smallmat<- bigdatamat[101:200,3:5]
      # creates small mat as the indicated subset with columns 3 to 5, rows 101 to 200
  smallmat<- bigdatamat[101:200,c(2,4)]
      # takes columns 2 and 4 of bigmat
  smallmat<-bigdatamat[bigdatamat[,1]>20171231,]
      # takes rows with date after 12/13/2017
  ```

  Can also use multiple conditions with the & (and) , | (or), and ! (not) signs. Do help(“Logic”) to see more

  ```
  retmat[retmat[,1]>20171231 & retmat[,1]>0,]          #
       selects all periors past 20171231 where the first return is >0.
  ```

- Building blocs

  ```
  1:5           # a sequence of numbers from 1 to 5
  c(1,5)        # the numbers 1 and 5
  c(vector1,vector2)         # concatenates two vectors into 1
  cbind(mat1,mat2)           # joins 2 matrices next to each other, column bind
  rbind(mat1,mat2)           # joins 2 matrices with mat1 on top of mat2, row bind
  length(vector)             # gives the length
  dim(matrix)                # gives the dimension
  ```

Basic loop

```
for (i in 4:10){
    sales[i] <- price[i]*quantity[i]
    }
```

But loops can be avoided very often: sales<-price*quantity #done

The simple * in R is not an inner product but the Hadamard (element-wise product)

The inner product of vectors or matrices: a %*% b

- prod and sum commands give the product or sums of elements of a vector.

  ```
  prod(1+ myret)-1                    # compound returns contained in vector myret
  ```

- Plots

  ```
  plot(x,y,pch=”O” ,xlab=”Xvalue”,ylab=”yvalue”)
      #pch allows you to choose the symbol.
  ```

  Other useful qualifiers:

  ```
  plot(ret1,ret2,  xlim= c(low,high),   ylim=      ) # axes limits,
  points(x1,y2, col=”red”)            # adds more points to the previous graph with red color
  title(“This is my title”)           # adds title to the plot
  ```

  Do help(“par”) for graphics tweaks

  ```
  abline(h=max(rety))                 # adds horizontal line at chosen value
  abline(v=   )                       # adds vertical line
  abline(a=… ,  b= … )                # adds line with intercept a, slope b
  abline(lsfit(x,y))                  # adds regression line to scatter plot
  ```

- Histograms

  ```
  hist(rets, nclass=40, prob=T)
      # always see if you can increase the default number of bins to make the histogram
      look more realistic.
      # prob=T scales the histogram to be a density
  ```

- Descriptive stats

  ```
  mean(myret)
  sd(myret)          # standard deviation
  var(myret)         # variance
  cor(ret1,ret2)     # correlation
  ```

  If myret is a matrix of returns, not a single vector. Then var(myret) computes the covariance matrix. Need to use the apply command (we don’t loop)

- apply(retmatrix,2,fun)

  # computes the function fun of each column (2) or row (1) of a matrix. Result is a vector.   Ex: apply(retmatrix[retmatrix$date<20130000,2:1584],2,mean)