QAM homework 1

Sumeng Wang

April 12, 2019

Collabration: Xiahao Wang

Question 1

After downloading data from CRSP, first we want to clean the data.
1. Universe of the stocks: As instructed, we subset out the stocks with Share Code 10, 11, and exchange code 1, 2, or 3.
2. Remove the NA data in the PRC column. The rationale behind this is that if PRC is missing, we will not get a market value, hence we will not use this in our market portfolio. Therefore it is safe to simply delete these data with missing PRC. Then we replace the missing value of DLRET and RET with 0. The reason behind this is that we use

$$cum\_div\_ret = (1 + ret)(1 + dlret) - 1$$

to calculate the cum-dividend return. If DLRET = 0, then $cum\_div\_ret = ret$. It is also the same the other way around.
3. Next we calculate the market cap by multiplying the absolute value of PRC and SHROUT. the negative PRC indicates this is an average of bid/ask price that day. Then we shift the market value down by 1 because the market portfolio is constructed using time t's return and time $t-1$'s weight. After this, the first entry of each PERMNO will get an NA. Again, we replace it by 0 because this indicates 0 weight in the market portfolio.
4. Then for each PERMNO, we can use "weighted.mean" function to get the value-weighted market portfolio.

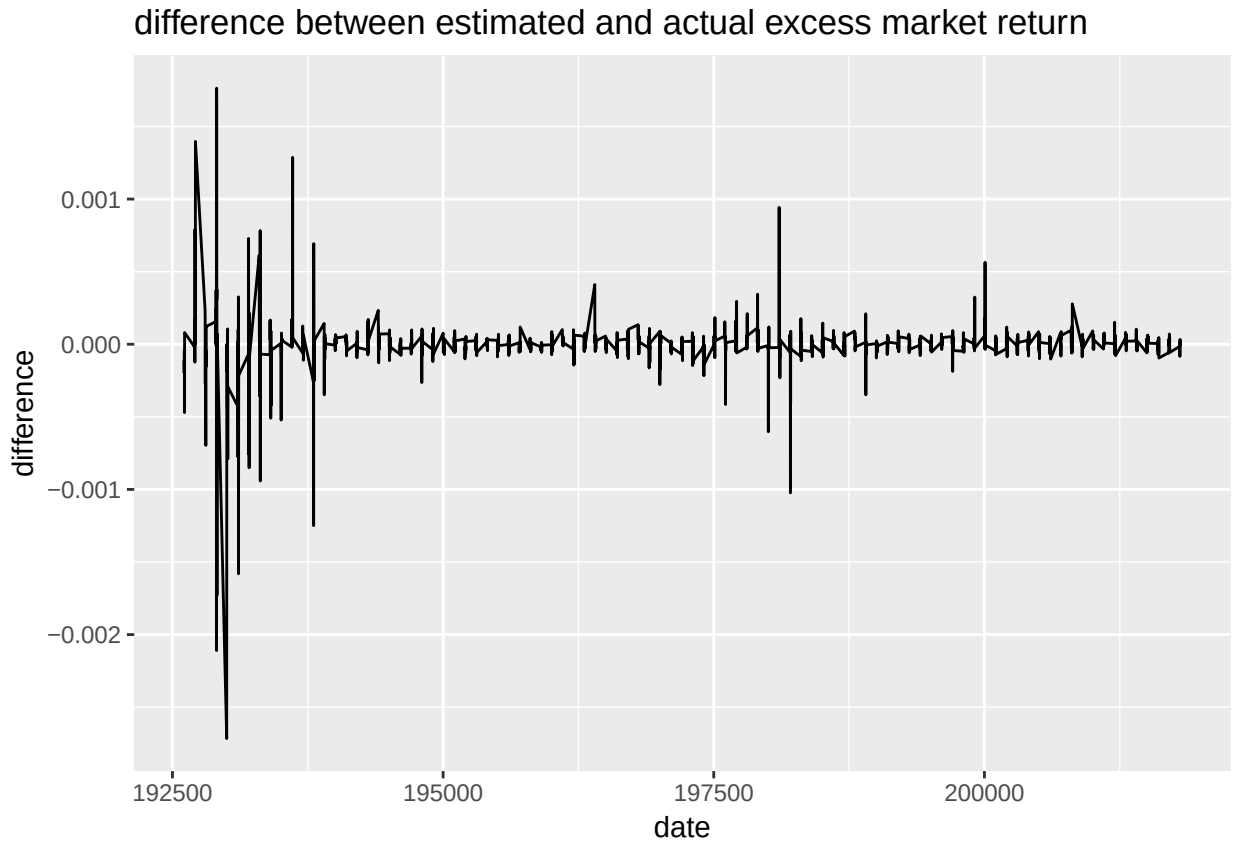
Question 2

Table 1: moments of market portfolio

Rm-Rf	estimated	actual
annualized mean	0.078000603783832	0.0780897297297297
annualized stdev	0.184697813056134	0.184772820947521
annualized SR	0.422314712303204	0.422625629295927
skewness	0.18061611295486	0.184006632942548
excess kurtosis	4.81263899984385	4.83095038359811

All the stats above are calculated using R's built in function.

Question 3



The correlation between two time series is 0.9999931. And the max absolute difference is 0.0027162. To explain the difference in two series, my thought is that the error occured during the data mining process. As we can see from the above figure, the differences between two series are larger at earlier times than those at later times. It is reasonable to guess that back in 1930s, 1940s, when we did not have computers to record all these data. Some data were missing or incorrect, which caused the inconsistency between my estimation and the one on French's website.