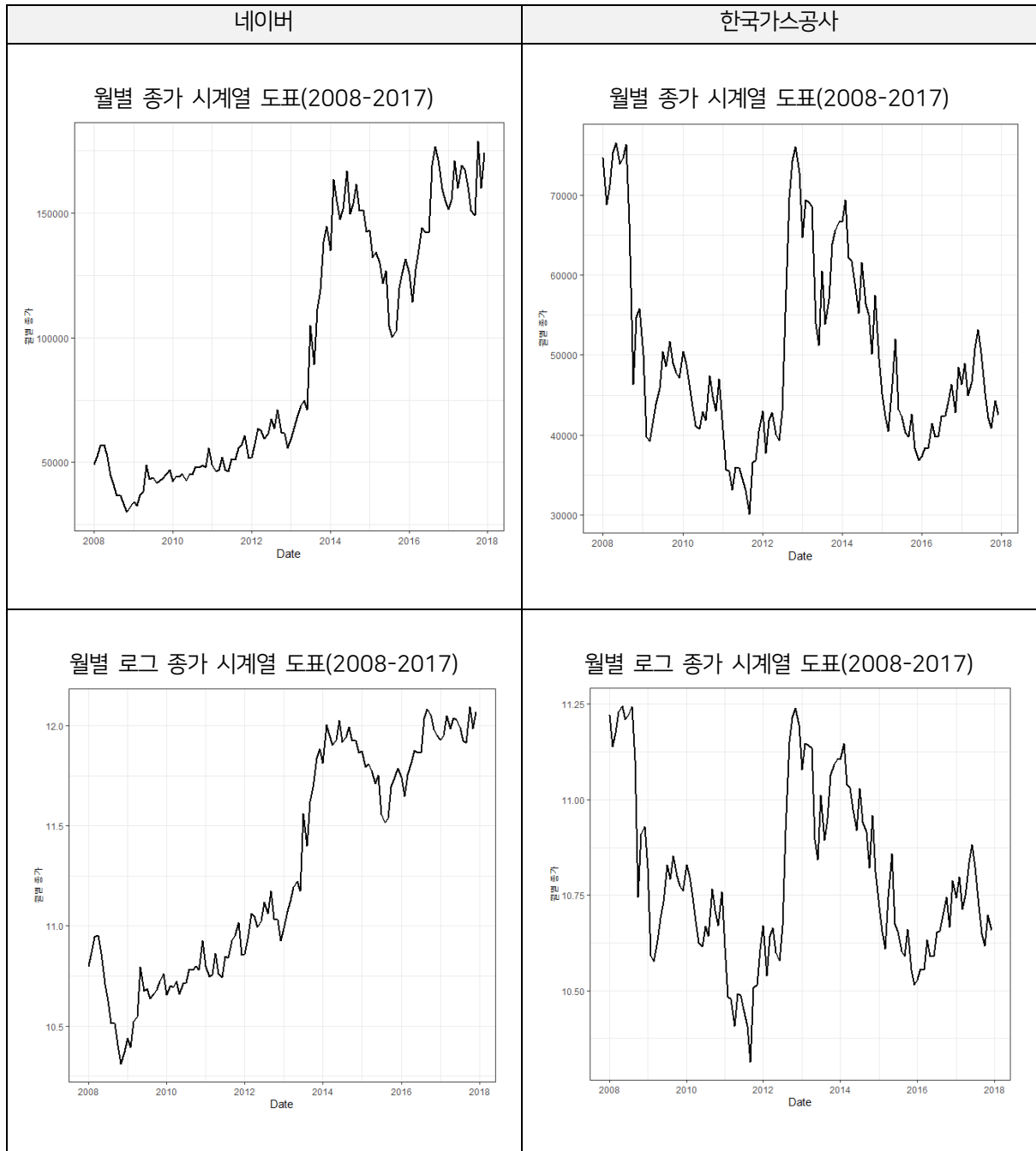


이론통계학2 - Project #5 Optimal Portfolio 계산/Application 개발 및 수익률 비교

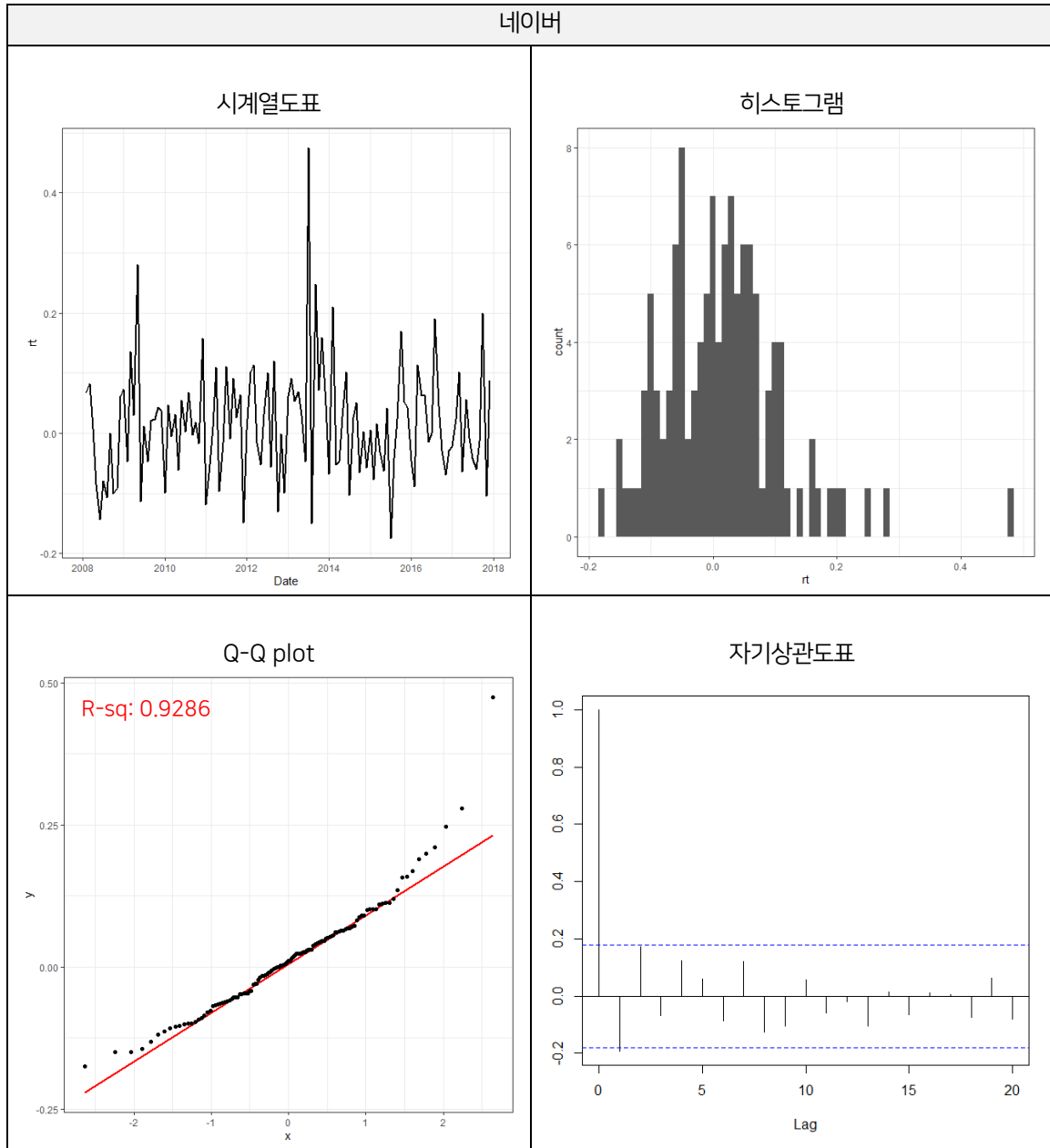
강아미, 고유정, 윤보인, 이혜린, 홍지원

Part 1) (주식투자 Portfolio 만들기)

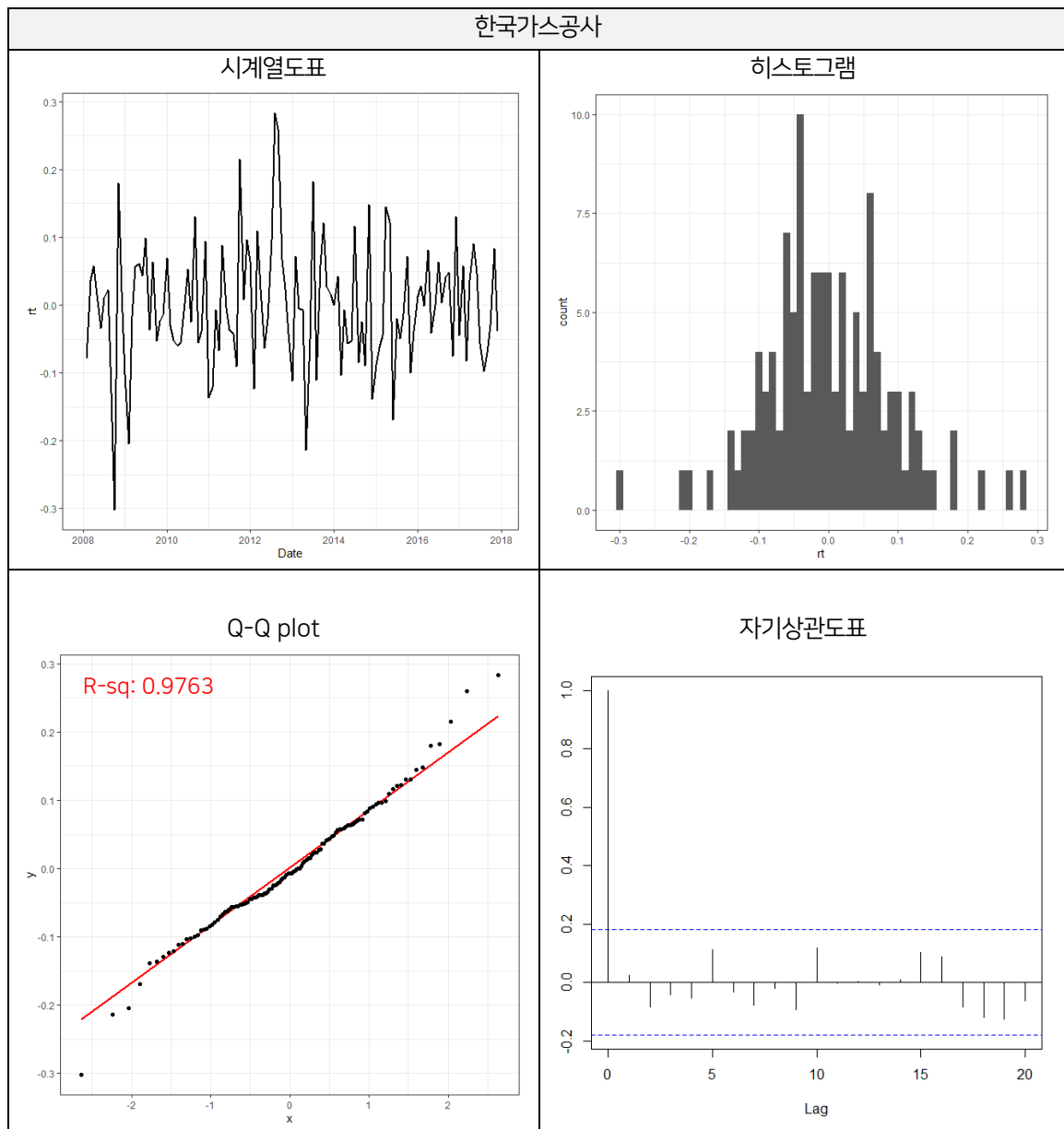
1) 각 종목의 2008-2017년 월별 종가 및 로그종가 $\{P_{it}\}, \{\ln P_{it}\}; t = 0, 1, \dots, (n + m)$ 에 대한 시계열 도표를 그리시오. $i = 1, 2$



2) 개별 종목의 2008-2017년 월별 수익률 $\{r_{it}\}$ 을 $r_{it} = \frac{P_{it} - P_{i(t-1)}}{P_{i(t-1)}}$, Portfolio: $P(\theta): (\theta_0, \theta_1, \theta_2); \sum_{j=0}^2 \theta_j = 1$ 의 월별 수익률 $r_{pt}(\theta)$ 을 $r_{pt}(\theta) = \sum_{i=0}^2 \theta_i \cdot r_{it}; r_{0t} = i_{12}(t); r_{it} = \frac{P_{it} - P_{i(t-1)}}{P_{i(t-1)}}, t = 1, \dots, (n+m)$ 로 정의할 때 r_{it} 의 자료의 시계열도표, Histogram, Q-Q Plot, 자기상관도표를 각각 그려보고 정규성(normality) 및 독립성 가정을 검토해 보시오. $i = 1, 2$



네이버의 월별 수익률 r_{1t} 을 이용하여 다양한 그림을 그려보았다. Q-Q plot의 선형성이 뚜렷하고 R-squared 값이 높으므로 r_{1t} 가 정규성 가정을 만족한다고 할 수 있다. 자기상관도표를 그려본 결과 대부분의 자기상관계수가 $2 \cdot se$ 경계 안에 존재하므로 r_{1t} 는 독립성 가정을 만족한다.



위의 그림은 한국가스공사의 월별 수익률 r_{2t} 을 이용하여 그린 그림이다. Q-Q plot의 선형성이 뚜렷하고 R-squared 값이 높으므로 r_{2t} 는 정규성 가정을 만족한다. 자기상관도표를 그려본 결과 대부분의 자기상관계수가 $2*se$ 경계 안에 존재하므로 r_{2t} 가 독립성 가정을 만족하는 것을 확인할 수 있다.

3) 학습자료(2008-2012) 및 검정자료(2013-2017)에 대해서 $\{r_{it}\}$ 의 평균과 분산($\hat{\mu}_i, \hat{\sigma}_i^2$) 및 공분산 행렬 $\Sigma = (\sigma_{ij}), i = 1, 2$ 을 각각 구하고 그 차이를 비교해 보시오.

Naver	Train	Test
$\hat{\mu}$	0.005648	0.02412
$\hat{\sigma}^2$	0.007240	0.01109

한국가스공사	Train	Test
$\hat{\mu}$	0.004646	-0.005507
$\hat{\sigma}^2$	0.010257	0.006872

학습자료 공분산행렬	
0.007240	0.0007237
0.0007237	0.010257

검정자료 공분산행렬	
0.01109	0.002115
0.002115	0.006872

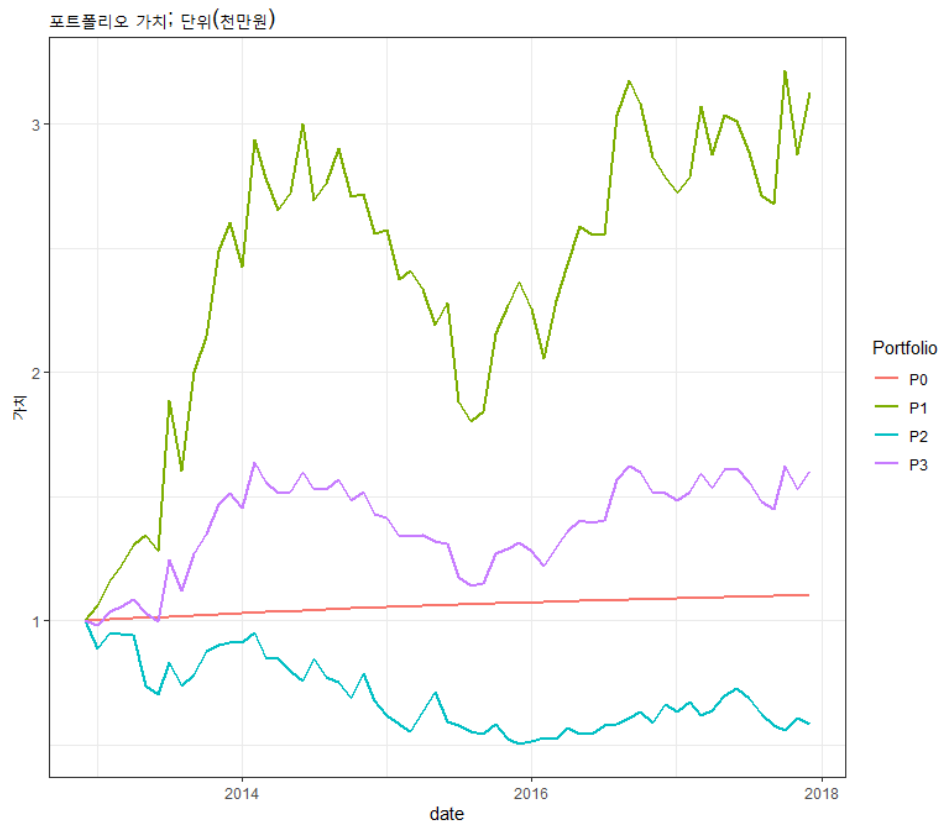
네이버의 경우 학습 자료(2008-2012)보다 검정 자료(2013-2017)에서 월별 수익률의 평균과 분산이 높았다. 이는 투자론의 기본인 "High Risk, High Return"에 부합하는 결과이다. 2013년부터 2017년까지, 네이버 주식의 평균 월별 수익률은 2퍼센트가 넘는데 이는 당시 연이자율인 2.8퍼센트와 비교했을 때 10배가량 높은 수치이다.

한국가스공사의 경우 검정 자료보다 학습 자료의 월별 수익률 평균과 분산이 높았다. 2013년부터 2017년 사이 한국가스공사의 월별 수익률 평균은 -0.55%였는데 이는 해당 기간동안 주가가 하락했음을 의미한다. 주식 투자의 기회비용인 정기 예금의 연이자율을 고려할 때 실질적 손실은 더욱 클 것으로 생각된다.

위의 표를 통해 학습 자료의 공분산보다 검정 자료의 공분산이 높은 것을 확인할 수 있다. 두 주가의 상관관계는 위험 분산을 위한 주식 포트폴리오 구성 시 고려할 수 있는 정보다. 투자자의 성향에 따라 포트폴리오를 달리 구성하기 때문에 공분산의 크고 작음이 특정 포트폴리오의 좋고 나쁨을 결정하지는 않는다.

4) 2013년 1월초에 총 $V_0 = 1$ (예를들어 1천만원)을 주식에 각각 $(\theta_0, \theta_1, \theta_2)$; $\sum_{j=0}^2 \theta_j = 1$ 비율로 분산투자했을 때 미래 t 개월 후 각 Portfolio의 가치(Value) $\{V_t\}$ 의 시계열도표를 동일한 그래프에 서로 적절히 겹쳐서 그려보고 각 Portfolio의 수익률의 장단점을 상호 비교 설명하시오. (단 $t = 1, \dots, m$; 2013년 1월-2017년 12월)

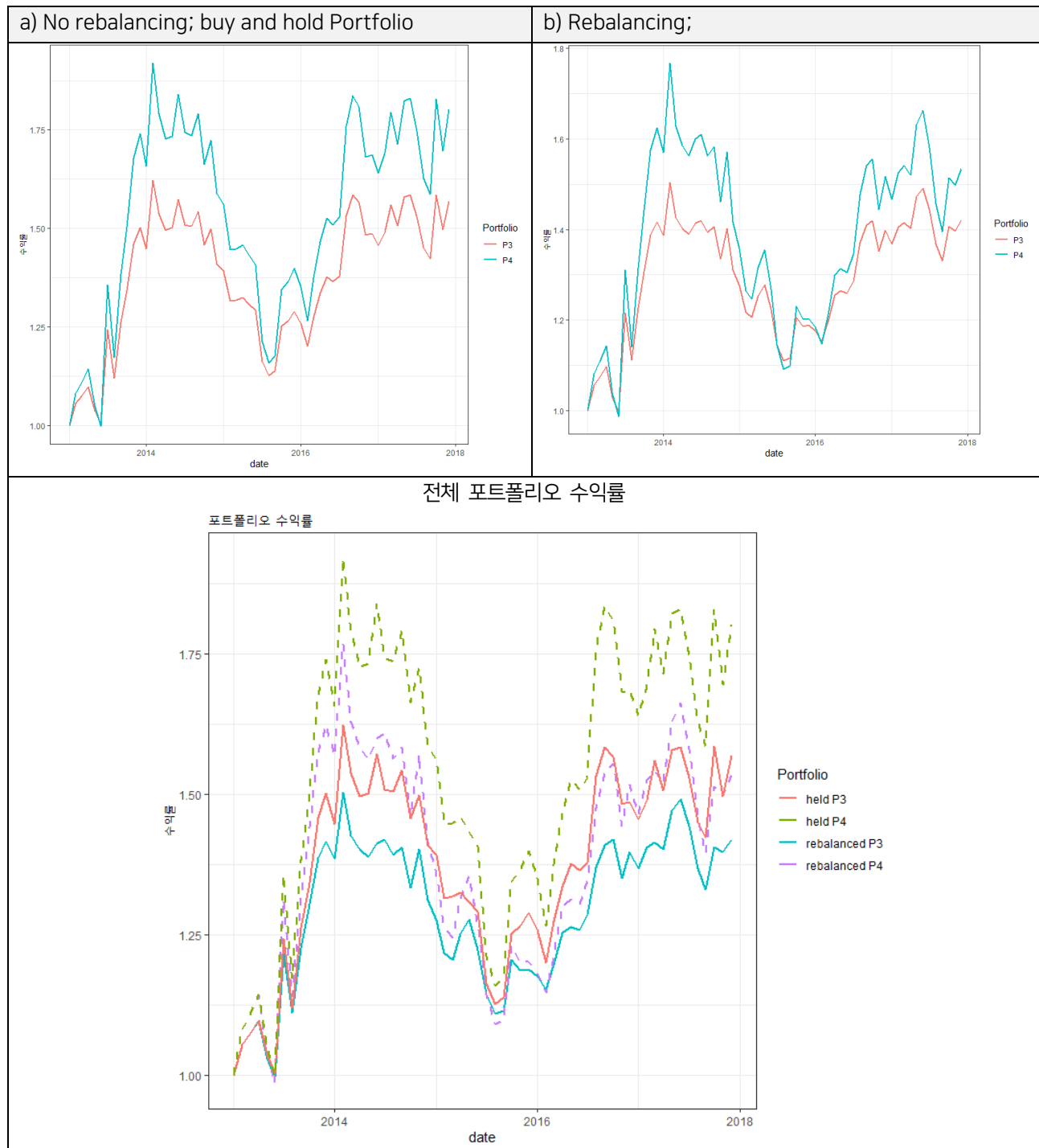
$$\begin{aligned} P_0: (\theta_0, \theta_1, \theta_2) &= (1, 0, 0) \\ P_1: (\theta_0, \theta_1, \theta_2) &= (0, 1, 0) \\ P_2: (\theta_0, \theta_1, \theta_2) &= (0, 0, 1) \\ P_3: (\theta_0, \theta_1, \theta_2) &= (1/3, 1/3, 1/3) \end{aligned}$$



P0은 자산 전액을 정기 예금에 투자한 포트폴리오이다. 손해볼 확률이 0에 가깝다는 장점이 있지만 그만큼 수익률이 낮다. P1은 자산 전액을 네이버에 투자한 포트폴리오로 가장 높은 수익률을 가진다. P2는 한국가스공사에 전액을 투자한 포트폴리오로 가장 낮은 수익률을 보이며, 해당 기간 포트폴리오 가치가 1000만원 이하로 떨어져 원금 손실을 초래하였다. P3는 정기 예금, 네이버, 한국가스공사에 1/3씩 투자한 포트폴리오로 두번째로 높은 수익률을 보인다. 분산 투자로 인해 수익률과 위험성이 모두 감소한 것을 확인할 수 있다.

5) 아래 두 Portfolio에 대해서 a) buy & hold 와 b) rebalancing 경우의 수익률 V_t^* , V_t 의 시계열도표를 서로 겹쳐서 그려보고 그 차이를 설명하시오. (단, $V_0 = 1$; $t = 1, \dots, m$; 2013년 1월 – 2017년 12월)

$$P_3: (\theta_0, \theta_1, \theta_2) = (1/3, 1/3, 1/3), P_4: (\theta_0, \theta_1, \theta_2) = (0, 1/2, 1/2)$$



위 그림을 통해 세가지 종목(정기예금, 네이버, 한국가스공사)에 투자한 P3 포트폴리오가 두가지 종목에 분산 투자한 P4 포트폴리오보다 수익률이 낮은 것을 확인할 수 있다. 더 많은 종목에 투자할수록 일반적으로 위험성(분산)이 작으며 수익률 또한 낮다.

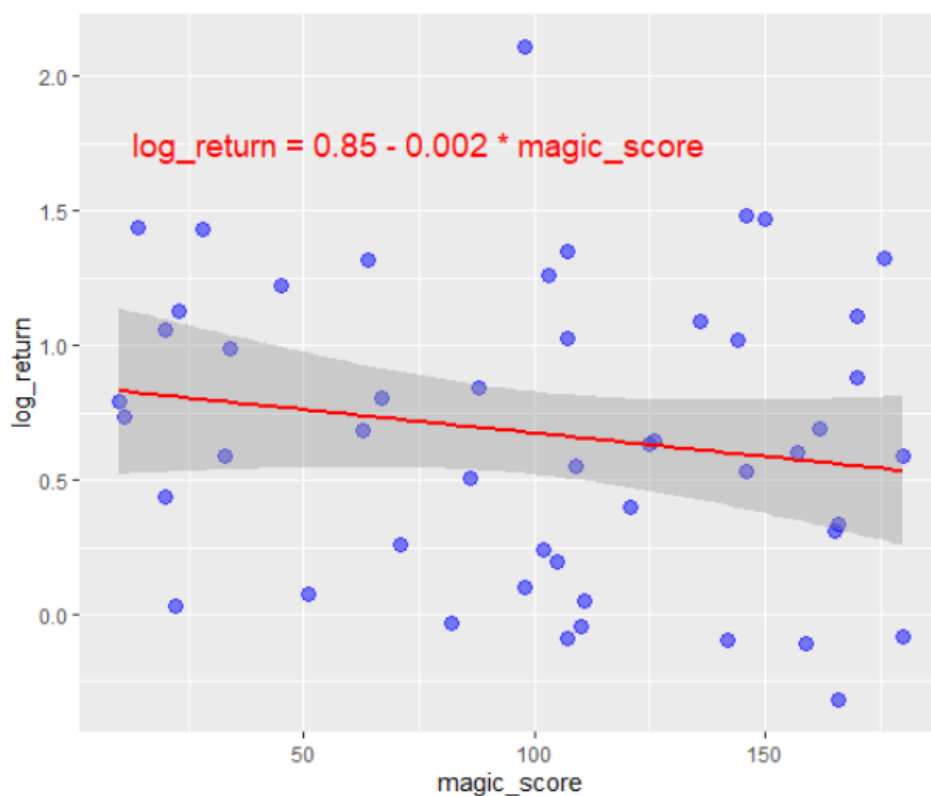
재조정된 포트폴리오(Rebalanced)의 분산이 투자 비율을 고정한 포트폴리오(Held)의 분산보다 작은 것을 확인할 수 있다. 전자의 경우 특정 종목에 자산이 치중되지 않게 지속적으로 조정함으로써 위험성을 감소시킨 것이다. 이는 비교적 낮은 수익률을 발생시켰다(Low risk, low return).

Part 2) (재무변수를 이용한 주가수익률 예측)

Data : 주식시장을 이기는 작은책 (The Little Book that Beats the Market), Joel Greenblatt (2006), 시공사, p 201-202 ; KOSPI 상장사 중 그린블랏의 마술공식(Magic Formula)점수 상위 50개 종목의 2005년 주가수익률 , 재무변수 (PER, ROA, Total Asset)자료

PER (주가수익비율)	PER이 낮으면 기업이 저평가되었다는 뜻. 즉 PER이 낮을수록 좋다.
ROA (총자산수익률)	기업의 이익발생능력의 의미. 즉 ROA가 높을수록 좋다.

1) 수익률과 그린블랏의 매직점수 (ROA순위+PER순위)의 산점도를 그리고 최적 예측방정식을 구하시오. 또한 두 변수의 상관계수 및 순위상관계수를 구하고 그 의미를 설명하시오.

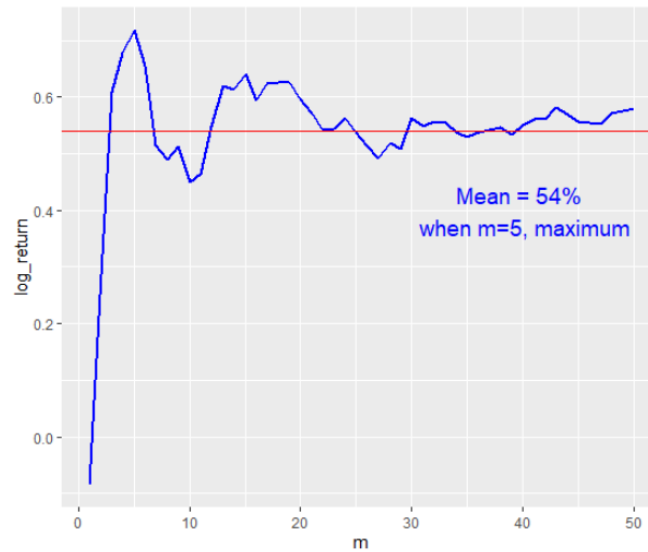


Magic score는 로그수익률에 아주 약한 음의 영향을 주고 있음을 알 수 있다. 이는 상관계수와 순위상관계수 값이 음수가 나온 것과 같은 맥락이다. 또한 상관계수와 순위상관계수를 통해 magic score와 로그수익률은 음의 선형, 비선형 관계를 갖고 있음을 알 수 있다(순위상관계수로 비선형관계 파악 가능). 그리고 변수 변환 없이 구해본 최적 예측방정식은 다음과 같다.

	Regression	Adj R squared
변수 선택 전	$\log(1+\text{return}) = 0.27 + 0.026\text{ROA} + 0.037\text{PER} + 0.00005 \text{ asset}$	0.2058
변수 선택 후	$\log(1+\text{return}) = 0.2743 + 0.026\text{ROA} + 0.00005 \text{ asset}$	0.2073

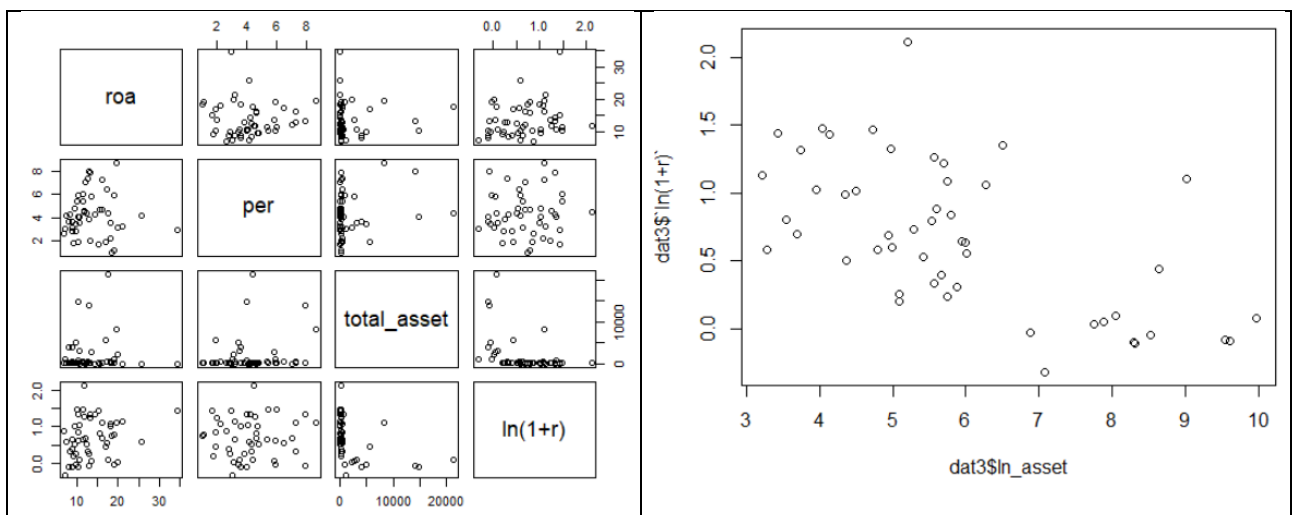
Variable selection (방향 = both) 를 통해 Adj R square를 아주 약간 올릴 수 있었지만 결과는 역시 좋지 않다.

2) 매직점수 상위 m 개의 종목에 $1/m$ 씩 분산투자한 포트폴리오의 수익률을 $m=1,2, \dots, 50$ 일 때 각각 구하여 이를 그래프로 나타내고 그 의미를 설명하시오.



분산투자한 포트폴리오의 로그수익률은 상위 5개 기업에 투자할 때 가장 크다. 로그수익률은 상위 5개 기업에 투자할 때까지 큰 속도로 증가하다가, 10개 기업에 투자할 때까지 하락한다. 그리고 그 이후로는 증가와 감소를 반복하며 평균 근처에서 머문다. 이를 통해 상위 몇 개 기업에 투자하느냐가 절대적으로 포트폴리오의 로그수익률을 결정한다고는 볼 수 없으며, 예측 시뮬레이션을 통해 적절한 기업 개수를 선택하는 것이 옳바르다는 것을 알 수 있다. 또한 분산 투자 후의 평균 로그수익률이 54%임을 통해 분산 투자는 효과적이라는 것을 알 수 있다.

3) 재무변수 (PER, ROA, Total Asset)들을 적절히 변환하여 로그수익률 ($\ln(1+r)$)을 예측하는 최적 회귀모형을 찾고 이를 a) 에서 구한 결과와 서로 비교하여 새 예측공식의 의미를 설명하고 이들 결과의 의미를 주식시장의 효율적 시장가설과 연관하여 해석하시오.



변수 변환 전에, pairplot을 그려봤을 때, asset과 로그수익률의 plot에서 값이 상당히 구석으로 몰려있음을 볼 수 있었기에 asset에 로그변환을 취했다. 덕분에 두 변수간 선형성을 찾을 수 있었다. 그 후, 축차적으로 변수 변환을 진행하여 Adj R squared를 높여나갔다.

Step	Regression	Adj R squared
Asset 로그변환	$\log(1+\text{return}) = 1.42 + 0.014 \text{ ROA} + 0.03 \text{ PER} - 0.18 \log(\text{asset})$	0.3525
ROA 로그변환	$\log(1+\text{return}) = 0.98 + 0.25\log(\text{ROA}) + 0.027\text{PER} - 0.18 \log(\text{asset})$	0.3614
PER 제곱변환	$\log(1+\text{return}) = 1.07 + 0.24\log(\text{ROA}) + 0.004\text{PER}^2 - 0.18 \log(\text{asset})$	0.3690

1)에서의 최적 예측방정식과 3)에서의 최적 예측방정식을 비교해보면 다음과 같다.

	Regression	Adj R squared
변수 변환 전	$\log(1+\text{return}) = 0.2743 + 0.026\text{ROA} + 0.00005 \text{ asset}$	0.2073
변수 변환 후	$\log(1+\text{return}) = 1.07 + 0.24\log(\text{ROA}) + 0.004\text{PER}^2 - 0.18 \log(\text{asset})$	0.3690

Adj R squared가 약 0.16 상승한 것을 통해 변수 변환이 적절했음을 알 수 있다. 또한 변수 변환 후, ROA와 asset이 로그수익률에 미치는 영향력은 높아진 것을 알 수 있다. (참고로 변수 변환 후 변수 선택 과정을 거쳤으나 결과가 좋지 않았다.)

효율적 시장가설이란 모든 시장참여자가 모든 정보를 가지고 있을 때 자산가격이 균형에 도달한다는 가설이다. 여기서 정보는 마지막 시점에 가격에 반영되며, 정보와 관련한 세가지 시점에 따라 강형, 준강형, 약형 효율적 시장가설로 나눌 수 있다.

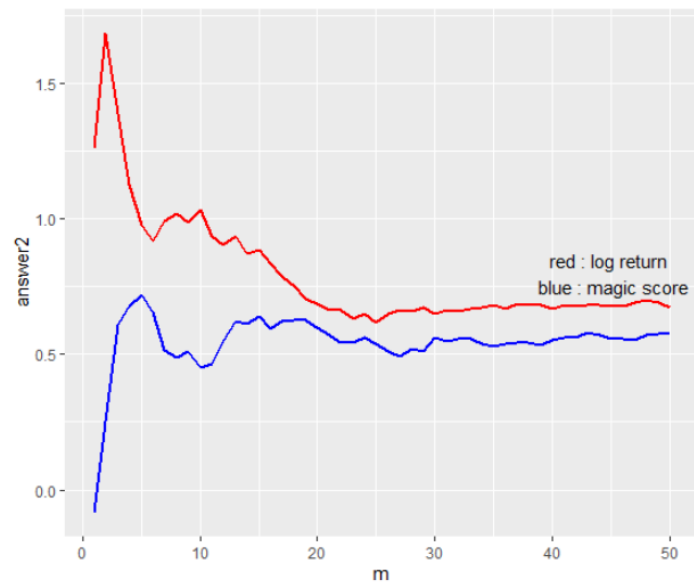
시점1. 정보가 발생하는 때 / 시점2. 정보가 공개되는 때 / 시점3. 모든 시장참여자가 정보를 획득하는 때

강형 효율적 시장가설	<ul style="list-style-type: none"> 세 시점이 모두 일치 (정보가 발생하자마자 가격에 반영) 비대칭 정보가 존재할 수 없음 시장참여자는 어떤 방법을 쓰더라도 시장수익률을 뛰어넘을 수 없음
준강형 효율적 시장가설	<ul style="list-style-type: none"> 세 시점 중 뒤 두 시점이 일치 (정보 공개 시점 = 정보 획득 시점) 정보가 발생한 뒤, 공개될 때 가격에 반영 시장참여자는 남들이 알지 못하는 정보를 가지고 있을 때에만 시장수익률을 뛰어넘을 수 있음 정보의 비대칭성이 해결되면 시장이 효율성을 달성
약형 효율적 시장가설	<ul style="list-style-type: none"> 세 시점이 전혀 일치하지 않음 정보가 발생하고, 공개된 뒤, 시간이 흐른 다음에야 가격에 반영 준강형과 마찬가지로 시장참여자는 남들이 모르는 정보를 알아야 시장수익률을 뛰어넘을 수 있음 준강형과 차이 : 준강형은 정보의 전파가 즉각적, 약형 효율적 시장은 상대적으로 느림. 그래서 남들이 알지 못하는 정보의 범위에 비대칭 정보뿐만 아니라 시간에 따라 이미 공개된 정보까지 포함될 수 있음. 그 결과 시장참여자는 정보공개의 출처와 시간적으로 가까운 위치를 차지하려는 전략을 선택

만약 강형 효율적 시장가설을 따른다면, 시장 참여자는 분산 투자 등의 어떠한 방법을 쓰더라도 시장수익률을 뛰어넘을 수가 없다.

그러나 약형 효율적 시장가설을 따른다면, 시장 참여자는 남들이 모르는 정보를 활용한다면 시장수익률을 넘을 수 있다. 여기서 남들이 모르는 정보란 최적 예측방정식을 통해 구한 분산 투자 후의 포트폴리오 로그수익률 등을 예로 들 수 있다.

4) 최적예측모형에서 구한 수익률의 추정값(fits) 순위 상위 m 개의 종목에 $1/m$ 씩 분산투자한 포트폴리오의 수익률을 $m=1,2, \dots, 50$ 일 때 각각 구하여 이를 b)에서구한 그래프와 겹쳐서 그리고 그 의미를 설명하시오.



로그수익률의 추정값을 가지고 순위를 매겨 상위 m 개의 기업에 분산투자 했을 때에는, 투자하는 기업의 개수가 2개 일 때 포트폴리오 수익률이 가장 높았다가 그 이후 감소하였다. 이 때도 매직 스코어를 이용했을 때와 마찬가지로, 시뮬레이션을 통해 몇 개의 기업에 분산투자할 때 가장 큰 수익을 낼 수 있는지를 미리 확인한 후 m 값을 정하는 것이 좋아 보인다.

또한 매직스코어보다 로그수익률의 추정값을 가지고 분산투자할 기업의 개수(m)를 정했을 때의 포트폴리오 수익률이 더 높았다. 이를 통해 추정된 로그수익률 예측방정식이 잘 추정되었음을 알 수 있다.

<part3> Optimal Portfolio 찾는 Application 개발 및 최적 포트폴리오 의미 해석

- 1) Optimal Portfolio 구하는 Application 개발
- 2) Problem: 주어진 목표수익률에 대한 Sharpe Ratio를 극대화하는 최적 포트폴리오를 구하는 앱 개발.

Optimal Portfolio

Data

File input

Browse...

GASshiny.csv

Upload complete

Browse...

KAshiny.csv

Upload complete

투자 시작 기간

2013

초기 투자 금액(v_0 , 단위:천만원)

1

정기 예금 비중

0 0.2 1

목표수익률(%)

1 7 100

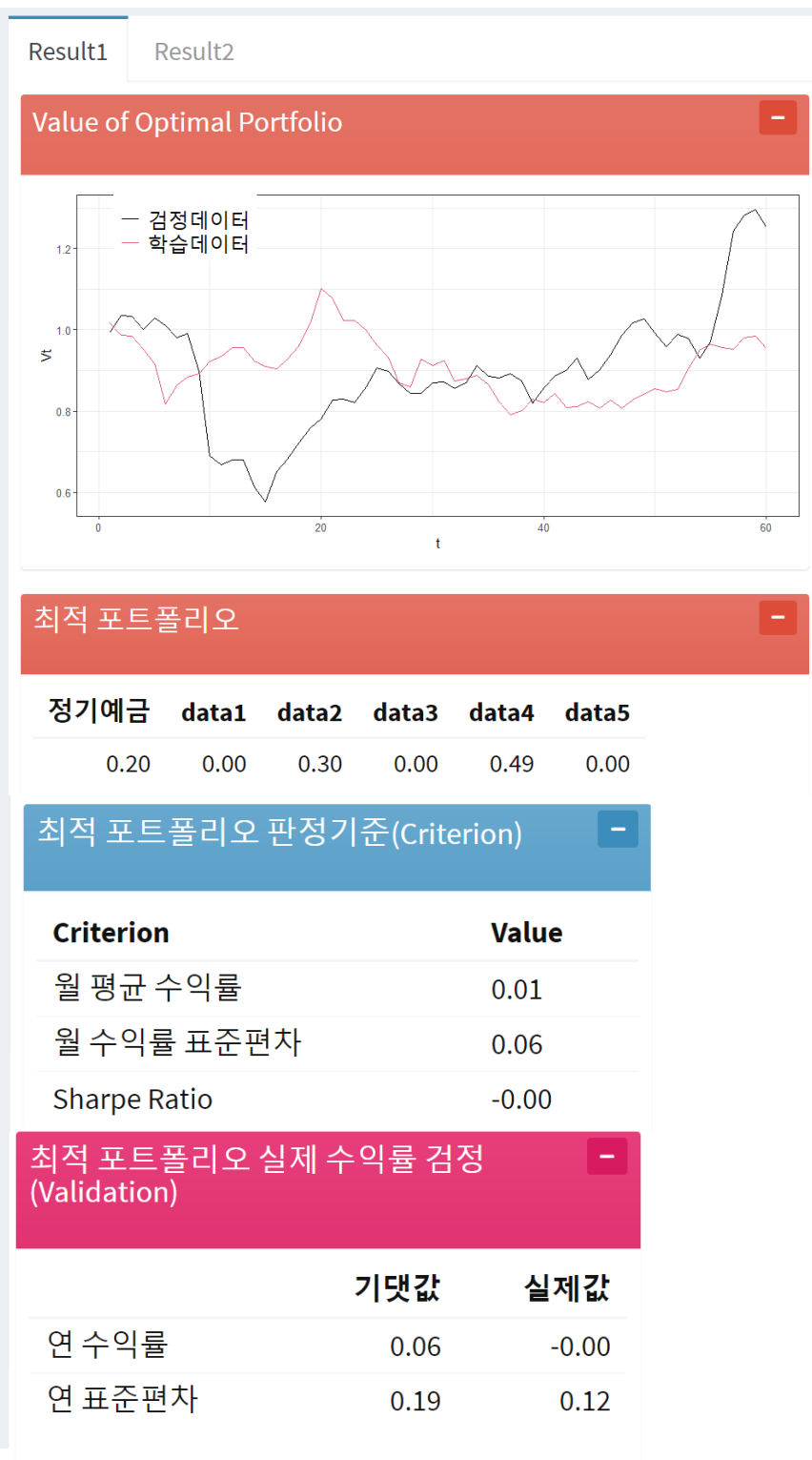
투자 기간(M, 단위:연)

5

표본 크기(N, 단위:연)

5

Go!



3) Optimal Portfolio Table 만들고 (예 : SolverTable 이용) 표 의미를 설명해보시오.
(정기예금 비율 0.2로 고정했을 때)

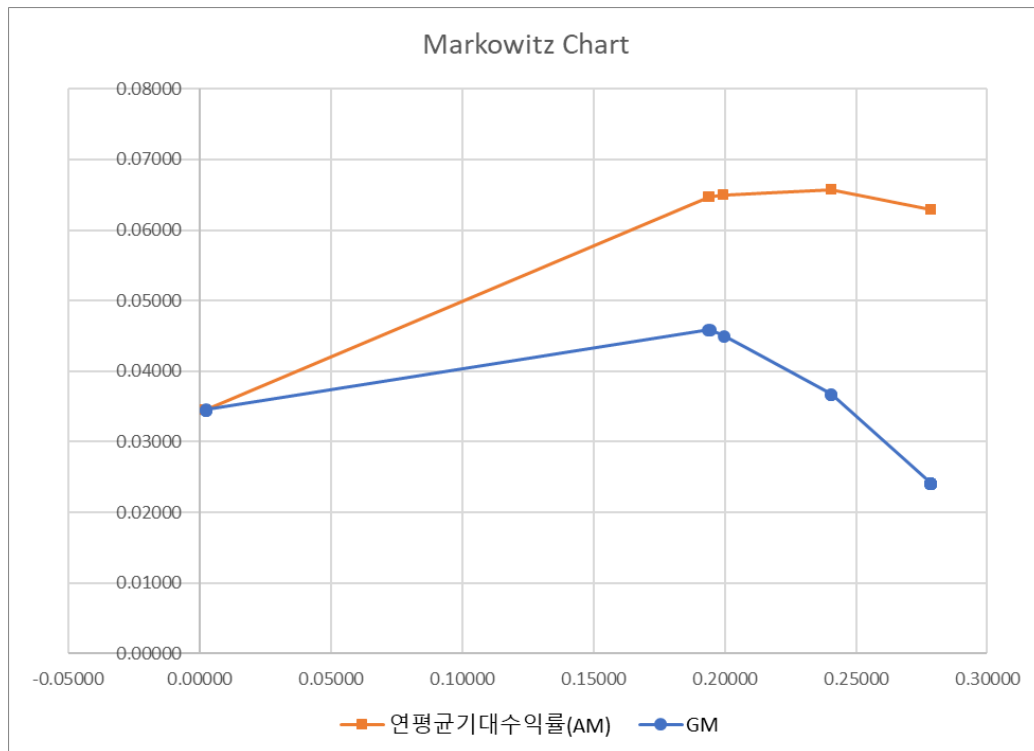
목표	정기 예금	삼성 전자	카카오	KT&G	한국 가스공사	네이버	MEAN	SD	Sharpe
1	0.9926	0.0013	0.0031	0	0.0002	0.0028	0.0029	0.0007	2.9855
2	0.9925	0.0013	0.0032	0	0.0002	0.0028	0.0029	0.0007	1.7846
3	0.9923	0.001	0.0036	0	0.0006	0.0026	0.0029	0.0007	0.5963
4	0.2	0	0.3049	0	0.4951	0	0.0054	0.0559	0.0377
5	0.2	0	0.2876	0	0.5124	0	0.0054	0.0561	0.0234
6	0.2	0	0.2165	0	0.5835	0	0.0054	0.0577	0.0094
7	0.2	0	0	0	0.8	0	0.0055	0.0694	-0.0026
8	0.2	0	0	0	0.8	0	0.0055	0.0694	-0.0139
9	0.2	0	0.8	0	0	0	0.0052	0.0804	-0.0245
10	0.2	0	0.8	0	0	0	0.0052	0.0804	-0.034
11	0.2	0	0.8	0	0	0	0.0052	0.0804	-0.0435
12	0.2	0	0.8	0	0	0	0.0052	0.0804	-0.0529
13	0.2	0	0.8	0	0	0	0.0052	0.0804	-0.0622
14	0.2	0	0.8	0	0	0	0.0052	0.0804	-0.0714
15	0.2	0	0.8	0	0	0	0.0052	0.0804	-0.0805
16	0.2	0	0.8	0	0	0	0.0052	0.0804	-0.0896
17	0.2	0	0.8	0	0	0	0.0052	0.0804	-0.0986
18	0.2	0	0.8	0	0	0	0.0052	0.0804	-0.1076
19	0.2	0	0.8	0	0	0	0.0052	0.0804	-0.1164
20	0.2	0	0.8	0	0	0	0.0052	0.0804	-0.1252
21	0.2	0	0.8	0	0	0	0.0052	0.0804	-0.134
22	0.2	0	0.8	0	0	0	0.0052	0.0804	-0.1426
23	0.2	0	0.8	0	0	0	0.0052	0.0804	-0.1512
24	0.2	0	0.8	0	0	0	0.0052	0.0804	-0.1598
25	0.2	0	0.8	0	0	0	0.0052	0.0804	-0.1683

목표수익률에 따라 Sharp ratio를 극대화하는 optimal portfolio를 구한 표는 위와 같다.

목표 수익률이 3%이하일 때는 변동이 작아 수익률이 안정적인 정기예금 위주로 투자하는 것이 최적인 포트폴리오라는 결과가 나왔고, 목표수익률이 4~6% 인 경우 카카오와 한국가스공사에 분산 투자하는 것이 최적이었다. 수익률이 9% 이상일 때는 정기예금 비중을 제외하고 카카오에 투자하는 것이 가장 최적인 결과가 나왔다. KT&G의 경우 투자율이 0%로 유지되는 것을 확인할 수 있다.

샤프 지수는 위험자산에 투자함으로써 얻은 초과수익의 정도를 나타내는 지표로 샤프 지수가 높을수록 좋은 펀드라고 볼 수 있다. Sharpe Ratio가 가장 높은 경우는 목표수익률이 1%일때며 Sharpe Ratio는 2.98이다. 표준편차 또한 제일 낮은 것으로 보아 최적의 포트폴리오 값이라고 볼 수 있다. 반면에 목표수익률이 7%이상이면 카카이를 제외한 나머지 주식들의 가중치가 0에 수렴하며 샤프지수는 마이너스 값을 띄게 되고 표준편차가 급격하게 커지는 것을 알 수 있다.

3) Markowitz Mean-Variance Chart 그리고 그래프 의미 해석



AM : 연평균 기대수익률

GM : 연평균 기대수익률 - (기대 연 표준편차)² / 2

SD : 기대 연 표준편차

Efficient Frontier 마코위츠의 효율적 경계선은 주어진 데이터를 총동원해서 만들 수 있는 모든 포트폴리오 중에서 가장 효율적인 포트폴리오를 의미한다. x축은 risk, y축은 expected return을 나타내며 이 둘의 상관관계를 보여주고 있다.

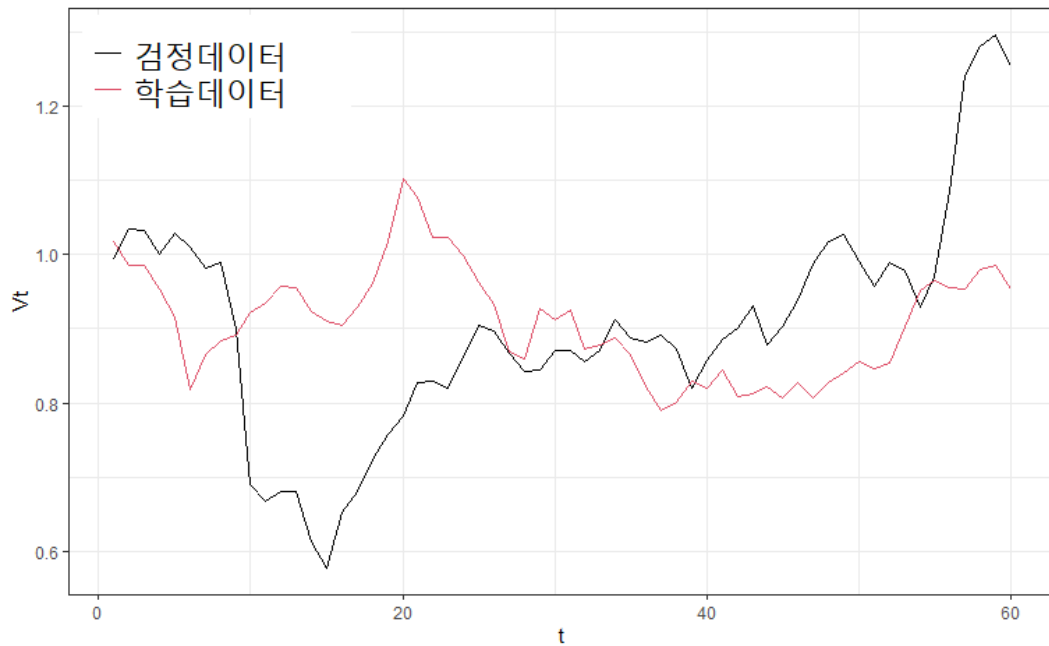
위 그래프에서는 AM이 GM보다 크게 나오는 경향을 띄며 수익률을 높게 잡을수록 작아진다. 즉, 변동성이 커지면서 낮아지는 경향을 보인다. GM을 사용한 포트폴리오는 AM을 사용했을 때에 비해 낮은 risk와 높은 기대수익률을 보여준다. 따라서 GM을 사용한 포트폴리오가 더 효율적이라고 볼 수 있다.

4) 1)에서 구한 최적포트폴리오 학습자료에서의 기대수익률과 검정자료에서의 실제수익률을 서로 비교해보고 차이가 있을 경우 그 이유 및 의미를 설명 해보시오.

학습자료(2008-2012)		검정자료(2013-2017)	
연 평균 수익률	0.0636	연 평균 수익률	-0.0002
연 평균 표준편차	0.1937	연 평균 표준편차	0.1241

수익률을 7%로 했을 때의 최적 포트폴리오 결과는 정기예금 정기예금 0.2%, 한국가스공사 0.3, 한국가스공사에 0.49% 투자하는 것이다. 학습자료에서의 기대수익률과 검정자료에서의 실제수익률은 차이가 나게되며 전반적으로 학습 자료의 기대수익률이 검정자료의 실제수익률 보다 높은 경향이 보이고, 특히 위 표의 결과처럼 수익률을 7%로 했을 때, 기대수익률에 크게 못 미치는 수익률을 냈다. 학습자료에서 나온 과거 5년간의 평균 수익률을 갖고 검정데이터의 5년치 수익률을 예측하는데 이는 과거의 평균 수익률과 표준편차가 미래 5년에도 반복됐음을 가정한 것이다. 주가의 표준편차는 과거와 미래가 동일할 수 없다. 과거 데이터를 갖고 수익률 등락을 정확히 예측할 수 없기 때문에 실제 수익률과 기대수익률은 차이가 나게 불가피하다.

5) 1)에서 그린 검정자료에서 2013년 1월초 초기 자산을 최적포트폴리오에 투자했을 때 t 개월 후 최적 포트폴리오의 가치의 시계열도표와 학습자료에서 구한 투자한 지 t 개월 후의 최적 포트폴리오의 가치를 겹쳐서 그려보고 두 시계열도표의 차이가 있는 경우 그 이유를 설명 해보시오.



검정자료의 t 개월 후 최적 포트폴리오 가치: $V_t = V_0 \prod_{s=1}^t (1 + r_{ps}^*) ; t = 0, 1, \dots, m$

학습자료의 t 개월 후 최적 포트폴리오 가치: $V_t = V_0 \prod_{s=1}^t (1 + r_{ps}) ; t = 0, 1, \dots, n$

검정자료의 r_{ps}^* 값과 학습자료의 r_{ps} 값이 다르기 때문에 두 시계열 도표에 차이가 난다. 전반적으로 검정데이터의 최적 포트폴리오 가치가 학습데이터의 최적 포트폴리오 가치보다 많이 높은 형태를 띈다. 이는, d 와 마찬가지로 과거의 r_{ps} 값과 검정데이터의 실제 r_{ps}^* 값은 똑같을 수 없고 실제 r_{ps}^* 값의 등락의 변동을 정확히 예측할 수 없다. 따라서, t 가 커질수록 검정데이터와 학습데이터의 포트폴리오 가치는 차이가 증가하는 추세를 보인다.