Math Capstone PBL (Data Analysis) – Project #1

CEO Characteristics and Firm R&D Spending

경제금융학부 2015024284 이재선

파이낸스경영학과 2017004093 양준우

목 차

제 1 장 서론

제 1절 연구 배경 및 목적

연구개발(R&D)에 돈을 쓰는 것은 기업의 최고 경영자(CEO)가 내릴 수 있는 근본적인 투자 결정 중 하나이다. 신제품, 프로세스 또는 기술 개발에 대한 기업 투자는 미래의 경쟁 우위 확보 및 생산성 증진의 원동력이 된다. 따라서 기업 R&D 지출의 증가 또는 감소와 관련된 요인에 대한 연구는 오래전부터 활발하게 진행되어 왔다. 구체적으로 이상석(2012)은 중견기업 CEO의 특성이 R&D 투자에 미치는 영향에 대한 연구를 진행하였다. 그는 기존 선행 연구와 달리 연구 대상을 중견기업으로 한정함으로써 CEO의 특성이 R&D 투자에 미치는 영향을 더 잘 설명할 수 있었다. 이는 대기업의 경우 CEO의 의사결정이 총수나 직계 존·비속에 크게 영향을 받기 때문에 기업의 투자 결정에 CEO의 특성이 전적으로 반영된다고 보기 어렵기 때문이다. 한편, Serfling(2014)은 CEO의 나이가 기업의 위험감행행동(Risk-taking Behavior)과 성과에 미치는 영향에 관한 연구를 하였다. Serfling(2014)에 따르면 기업의 위험감행행동은 CEO의 나이가 많을수록 감소하는 경향을 보이며, 따라서 R&D에 덜 투자한다고 한다. Benmelech & Frydman(2015)은 군 복무 경험이 있는 CEO는 불확실한 선택을 하는 것을 꺼리며, 따라서 투자와 R&D 비중을 줄인다는 연구 결과를 내놓았다. 본 연구에서는 성별, 나이 등 CEO의 전반적인 특성이 기업의 R&D 투자와 어떤 상관성이 있는지 살펴보고자 한다. CEO의 특성 중 한 요인을 집중적으로 분석한 앞선 연구와 달리, 본 연구는 여러 요인에 대해 총체적으로 바라봄으로써 CEO의 특성이 R&D 투자와 어떻게 관련되어 있는지에 대하여 전반적으로 언급하는데 의의를 두고자 한다. 본 연구를 통하여 답하고자 하는 질문을 정리해보면 다음과 같다.

첫째, 기업의 R&D 투자와 관련 있는 기업의 재무적 특성은 무엇이 있는가?

둘째, CEO의 특성을 나타내는 변수는 R&D 투자에 대한 설명력이 없는가?

셋째, 설명력이 있다면, 어떤 항목이 상관성을 보이는가?

제 2절 연구 구성

본 연구의 구성은 다음과 같다. 제 1장의 연구 배경 및 목적에 이어, 제 2장에서는 자료를 소개하고, 분석을 진행하고자 한다. 우선 수집된 자료 내 변수를 소개하고, 기초적인 정보를

나타내고자 한다. 한편 자료 내 결측치가 존재하기 때문에 이를 적절히 처리해야 하는데, 이에 대한 과정도 상세히 기술하였다. 다음으로 각 변수별 기초통계량을 제시하였다. 변수의 기초통계량을 살펴보며 변수변환이 필요한 경우 이를 적절히 시행하였다. 이후 종속변수와 설명변수간 상관분석을 진행하였으며, 이 결과를 바탕으로 종속변수와 상관성을 찾을 수 없는 설명변수는 연구에서 제외하도록 하였다. 다음으로 설명변수간 상관분석을 통해 설명변수 사이의 독립성을 확보할 수 있는지 살펴보고자 하였다. 이를 통해 상관성이 뚜렷한 변수는 연구에서 제외하였다. 제 3장에서는 제 2장에서의 자료 분석 결과를 정리하고, 향후 분석 방향을 제시하였다.

제 2장 본론

제 1절 자료 소개

가. 자료 출처 및 소개

본 연구는 Wharton Research Data Services(WRDS) Compusata Data / Execucomp[1]에서 자료를 얻었다. 한양대학교 백남학술정보관 학술DB를 통하여 해당 서비스에 접근할 수 있었다. 2014년부터 2017년까지 기업의 재무적 성과와 CEO의 특성에 관한 총 7,409개의 패널 데이터를 얻었다. 한편, 모든 데이터를 분석에 사용하면 연도별 반복되는 기업의 데이터가 편향을 낳을 가능성이 존재한다. 따라서 본 연구에서는 비교적 최근인 2017년의 데이터를 가지고 분석을 진행하였다. 2017년 자료는 27개 변수로 총 1,726개이며, 결측치 처리 후 1,429개의 데이터를 분석에 사용하였다. 결측치 처리에 관한 내용은 뒤이어 자세히 설명하도록 하겠다.

한편, 데이터에 포함된 변수는 다음과 같다. 모든 변수는 '기업의 식별을 위한 변수', '기업의 재무적 특성을 나타내는 변수', 그리고 'CEO의 특성을 나타내는 변수'의 세 가지로 분류할 수 있다. 개별 변수에 대한 설명은 다음의 표로 제시하였다.

변수명	설 명
<기업 식별을 위한 변수>	
GVKEY	S&P Capital IQ/ Compustat/ Execucomp에서 사용되는 기업의 식별자이다.
SIC SIC2D SIC3D	기업의 주요 산업부문을 나타내는 4자리수 숫자이다. SIC2D, SIC3D는 각각 SIC의 첫 2자리, 3자리까지의 숫자를 의미하는데, 첫 두자리 숫자로 대분류가 가능하며, 세자리까지의 숫자로 중분류가 가능하다.
STATE	기업의 본사가 위치한 미국 주(州)를 의미한다.
<기업의 재무적 특성을 나타내는 변수>	
SIZE	기업의 총자산으로, 기업의 규모를 의미한다.
BM	시장가 대비 장부가 비율을 뜻한다. BM이 높은 것은 장부가격에 비하여 시장가격이 과소평가된 상황을 의미한다.
FCF	기업에 현금이 얼마나 순유입 되었는지를 나타내는 지표인 잉여현금흐름을 기업의 총자산으로 나눈 수치이다.
HHI	Herfindhal-Hirschman 지수이다. HHI는 0에서 1 사이의 값을 보이는데, HHI가

[1] https://wrds-www.wharton.upenn.edu/

	더 높을수록 기업이 더 경쟁적인 산업에서 활동하고 있음을 나타낸다.
OPPERF	기업의 영업성적을 총자산으로 나눈 수치이다.
LEVERAGE	기업의 레버리지의 시장가치를 기업 자산의 시가 평가액으로 나눈 수치이다.
TOBINSQ	TOBIN'S Q를 나타내는 변수로, 기업의 시장가치를 자본의 대체비용으로 나눈 값이다. 기업은 TOBIN'S Q를 투자에 대한 지표로 삼는데, Q값이 1보다 큰 경우 보유자산을 대체하는 데 드는 비용보다 시장가치가 크다는 의미이므로 투자를 통해 기업의 가치를 높이고자 한다.
RNDMISSING	재무제표 상 기업의 R&D 지출이 누락된 경우 1을 부여하는 DUMMY변수이다.
RNDRATIO	기업의 R&D지출을 총자산으로 나눈 값이다.
ROA	자산수익률을 의미한다. 기업이 자산을 얼마나 효율적으로 운용했는지를 나타낸다.
SALESGROWTH	전년대비 기업 판매량 증가율을 의미하는 변수이다.
DIVPAY	해당 회계연도에 배당금 지급이 이루어진 경우 1을 부여하는 DUMMY변수이다.
CASHRATIO	기업의 현금보유를 총자산으로 나눈 수치이다.
INTAN	무형성을 나타내는 지표로, 해당 값이 높을수록 기업의 무형자산 비중이 높다는 의미이다. 기업의 총자산에서 유형자산 가치를 뺀 값을 총자산으로 나눈 수치이다.
INVEST	기업의 투자비율을 의미한다. 투자액을 유형자산 가치로 나눈 수치이다.
<CEO 특성을 나타내는 변수>	
CEOAGE	CEO의 나이를 의미한다.
CEOCOMP	CEO의 총보수를 $1,000 단위로 나타낸 변수이다.
INSIDERCEO	CEO가 기업 내부에서 승진한 경우 1을 부여하는 DUMMY변수이다.
FEMALECEO	CEO의 성별이 여성인 경우 1을 부여하는 DUMMY변수이다.
CEOPAYSLICE	CEO의 보수를 기업 내 상위 5인의 보수를 더한 값으로 나눈 수치이다. 해당 값이 더 높을수록 CEO가 경영진 내에서 불균형적으로 많은 보수를 받고 있다는 의미이다.
CEOEQUITY	CEO의 기업 주식 보유를 나타낸다.
CEOTENURE	CEO의 재임기간을 연도 단위로 나타낸 수치이다.

<표1> 변수설명

나. 결측치 처리

자료분석에 앞서 누락된 데이터를 올바르게 처리하는 것은 매우 중요하다. 사회 현상을 연구하는 경우 자료에 결측치가 포함되는 경우가 빈번한데, 결측치 처리가 부적절하게 이루어질 시 심각한 오류를 일으킬 수 있다. 우선, 결측치는 귀무가설이 거짓일 때 이를 올바르게 기각할 확률을 나타내는 통계적 검정력(Statistical Power)을 감소시킨다. 또한, 결측치는 모수 추정에 있어

서 편향을 낳을 수 있다. 세 번째로 결측치는 표본의 대표성을 줄이며, 마지막으로 데이터 분석을 복잡하게 만든다.[2] 따라서 결측치 처리는 데이터 분석의 유효성과 맞닿아 있는 중요한 문제이며, 이를 적절한 방법에 따라 수행해야만 한다.

본 연구는 27가지 변수로 구성된 1,726개의 데이터를 활용하였다. 해당 자료도 결측치를 포함하고 있었는데, 변수별 결측치 개수는 다음과 같다.

변수명	결측치 개수(전체 자료 대비 비율)
<기업 식별을 위한 변수>	
GVKEY	0
SIC	0
SIC2D	0
SIC3D	0
STATE	56 (3.24%)
<기업의 재무적 특성을 나타내는 변수>	
SIZE	1 (0.06%)
BM	18 (1.04%)
FCF	104 (6.03%)
HHI	0
OPPERF	104 (6.03%)
LEVERAGE	36 (2.09%)
TOBINSQ	18 (1.04%)
RNDMISSING	0
RNDRATIO	779 (45.13%)
ROA	2 (0.12%)
SALESGROWTH	5 (0.29%)
DIVPAY	0
CASHRATIO	2 (0.12%)
INTAN	101 (5.85%)
INVEST	120 (6.95%)
<CEO 특성을 나타내는 변수>	
CEOAGE	1 (0.06%)
CEOCOMP	1 (0.06%)
INSIDERCEO	0
FEMALECEO	0

[2] Kang H. (2013). The prevention and handling of the missing data. *Korean journal of anesthesiology*, *64*(5), 402–406. https://doi.org/10.4097/kjae.2013.64.5.402

CEOPAYSLICE	266 (15.41%)
CEOEQUITY	29 (1.68%)
CEOTENURE	21 (1.22%)

<표2> 변수별 결측치 개수

본 연구는 우선적으로 변수별 5개 이하의 결측치를 가지는 데이터를 제거하였다. 따라서 SIZE, ROA, SALESGROWTH, CASHRATIO, CEOAGE, 그리고 CEOCOMP의 결측치는 모두 제거되어 총 1,719개의 데이터로 정리하였다. 다음으로 그 외 변수의 결측치를 산업부문별 평균값으로 추정하여 분석을 진행하고자 하였다. 이는 기업 데이터의 경우 기업 활동이 이루어지고 있는 산업부문에 따라 특징이 상이하게 나타날 수 있기 때문이다. 결측치를 전체 평균값이 아닌 산업별 평균값으로 나타냄으로써 산업부문에 따른 특징을 유지하고자 노력하였다. 우선 가장 소분류인 SIC별 평균을 결측치에 대입하였으며, SIC에 따른 평균을 알 수 없는 경우 순차적으로 SIC3D, SIC2D별 평균값을 대입하였다. 한편 이들 중 어느 것도 사용할 수 없는 경우 분석에서 제외하였는데, RNDRATIO 변수에서 282개의 데이터가 제외되어 총 1,437개의 데이터로 정리할 수 있었다. 이로서 모든 결측치를 적절하게 처리할 수 있었다.

한편, CEOCOMP의 값이 0인 값과 0.001인 값을 자료에서 제거하였다. 이는 CEO의 총보수가 각각 $0와 $1라는 의미인데, 현실성이 지나치게 떨어진다고 판단하여 분석에서 제외하고자 하였다. CEOCOMP 값이 0.001 다음 상위값은 0.637이었으며, 이는 어느정도 현실성을 가지기 때문에 제거하지 않고 분석을 진행하였다. 따라서 CEOCOMP가 0인 값 6개와 0.001인 값 2개가 제거되었으며, 총 1,429개의 데이터로 정리할 수 있었다.

제 2절 자료 분석

가. 기초통계량

본격적으로 자료를 분석하기에 앞서 데이터를 올바르게 가공해야 한다. 결측치를 적절한 방법으로 처리하는 한편, 변수가 분석에 용이하도록 정리해야 한다. 본 절에서는 변수의 기초통계량을 살펴보고, 데이터 가공이 필요한 경우 적절한 방법을 통하여 변수를 변환하는 과정을 서술하고자 한다. 본 연구에서는 로그변환, 제곱근변환, 그리고 ARCSINE변환을 통하여 변수를 분석에 알맞게 정리하였다. 이후 최종적으로 분석에 사용하는 변수의 기초통계량을 표를 통해 요약함으로써 이를 한 눈에 파악할 수 있도록 노력하였다. 한편 기업의 식별자를 나타내는 변수인 GVKEY

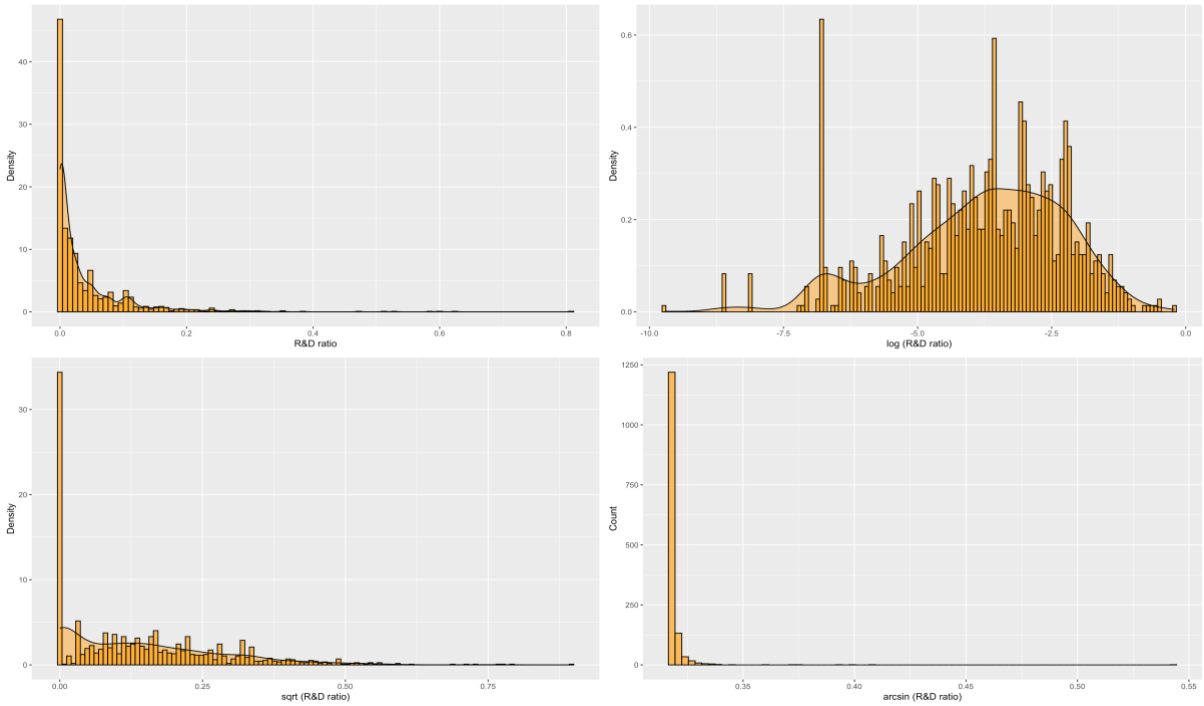
는 R&D 투자와 무관한 것이 자명하므로 분석에서 제외하였다. 산업부문을 나타내는 SIC, SIC3D, SIC2D의 경우 각각 326개, 226개, 65개로 분류가 되어있는데, 해당 변수를 추가적으로 소수의 범주로 대분류하기가 어려워 분석에서 제외하고자 하였다.

본 연구는 RNDRATIO를 종속변수로 설정하였다. 회귀분석 시 비율을 나타내는 변수를 포함하는 것은 허위상관(Spurious correlation) 등 오류를 발생시킬 가능성이 있다. 또한 잘못된 추론을 이끌어낼 수 있으며, 분석이 별다른 소득이 없이 끝날 수 있다.[3] 한편 회귀분석 시 종속변수의 정규성을 확인하는 과정은 중요하다. 이는 회귀분석의 오차항이 정규분포를 따른다는 가정을 하기 때문인데, 종속변수의 분포가 정규분포와 크게 다른 경우 문제가 될 수 있다. 따라서 본 연구에서는 RNDRATIO의 기초통계량 및 정규성을 확인한 후 정규분포 형태와 유사하도록 변환하고자 하였다. <그림1>에 나타난 것처럼 원 자료 내 RNDRATIO 변수는 왼쪽으로 크게 치우친 형태를 보였다. 왜도 또한 4.033으로 나타났으며, Shapiro-Wilk test 결과 정규분포를 따른다고 가정하기 어려웠다. 따라서 RNDRATIO에 대하여 적절한 변환을 할 필요가 있었다. 우리는 RNDRATIO에 대하여 로그변환, 제곱근변환, 그리고 ARCSINE변환을 고려하였는데, 각 변환에 따른 히스토그램의 변화는 <그림1>에서 확인할 수 있다. 로그변환 시 분포가 정규분포와 그나마 유사한 모습을 보였으나, RNDRATIO 변수의 값이 0인 경우 로그변환 시 -Inf로 변환되는 한계가 있었으며, 해당 수치가 393번(27.5%) 나타나 데이터를 제외하기 어려웠다. 한편 제곱근변환이나 ARCSINE변환으로는 정규분포와 유사하게 변환하기 어려웠다.

한편 RNDRATIO가 0인 집단과 0보다 큰 집단으로 분류하여 다른 설명변수와의 집단분석을 한 후 유의미한 차이가 보이지 않았을 때 0인 집단을 분석에서 제외하고 나머지를 로그변환을 하여 사용하는 방법도 고려해보았다. 그러나 ROA, INTAN, CEOAGE, CEOPAYSLICE, 그리고 CEOTENURE을 제외한 모든 변수에서 두 집단은 평균의 차이를 보였기 때문에 이들을 동질적인 집단으로 파악할 수 없었으며, RNDRATIO가 0인 집단을 제외할 수 없었다. 또한 RNDRATIO가 0인 집단을 제외할 시 분석의 모집단이 'R&D 분야에 투자를 하는 기업'으로 바뀌게 되어 본래의 연구 방향성이 틀어지게 된다. 따라서 본 연구에서는 추가적인 변환을 하지 않고 RNDRATIO를 종속변수로서 분석하기로 하였다.

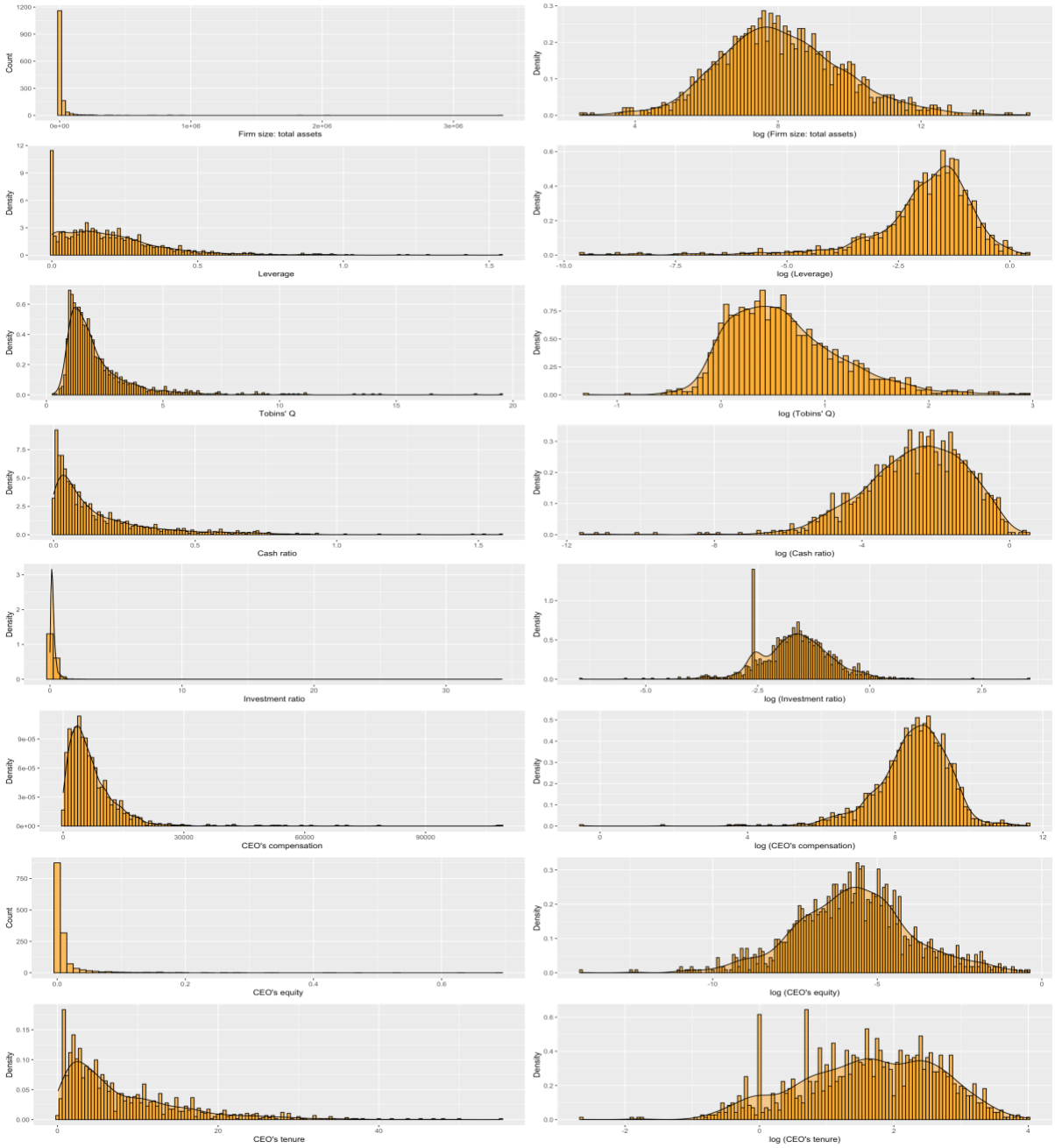
---

[3] Kronmal, R. (1993). Spurious Correlation and the Fallacy of the Ratio Standard Revisited. *Journal of the Royal Statistical Society. Series A (Statistics in Society), 156*(3), 379-392. doi:10.2307/2983064



<그림1> RNDRATIO 제곱근변환, 로그변환, ARCSINE변환 전후 히스토그램

비대칭적 분포를 보이는 독립변수에 대해서도 변환을 진행하였다. 독립변수가 지나치게 비대칭적인 분포를 가지는 경우 종속변수와의 상관성을 가지기 어려울 수 있기 때문이다. 우선 SIZE, LEVERAGE, TOBINSQ, CASHRATIO, INVEST, CEOCOMP, CEOEQUITY, 그리고 CEOTENURE 변수는 로그변환을 통해 비대칭성을 줄일 수 있었다. <그림2>에서 각 변수의 로그변환 전후 히스토그램 변화를 알 수 있으며, 변환 전후 기초통계량은 <표3>에 기술하였다. <그림2>와 <표3>을 통해 변환 후 비대칭성이 감소한 것을 파악할 수 있다.



<그림2> SIZE 등 8가지 변수의 로그변환 전후 히스토그램

변수명	평균	표준편차	Q1	Q3	최댓값	최솟값	첨도	왜도
SIZE	21997	129404.3	1034	9538	3345528	12	376.512	17.462
LOG(SIZE)	8.098	1.745	6.941	9.163	15.023	2.483	0.470	0.333
LEVERAGE	0.203	0.189	0.070	0.284	1.544	0	6.287	1.915
LOG(LEVERAGE)[4]	-1.951	1.266	-2.319	-1.217	0.435	-9.648	8.524	-2.336
TOBINSQ	2.200	1.724	1.236	2.494	19.549	0.278	24.311	3.963
LOG(TOBINSQ)	0.613	0.544	0.212	0.914	2.973	-1.281	1.105	0.869
CASHRATIO	0.159	0.191	0.033	0.213	1.580	0	8.335	2.393
LOG(CASHRATIO) [5]	-2.568	1.421	-3.418	-1.543	0.458	-11.645	3.454	-1.068
INVEST	0.282	0.963	0.116	0.308	34.132	0	1071.384	31.164
LOG(INVEST) [6]	-1.649	0.798	-2.129	-1.171	3.530	-6.430	3.174	-0.060
CEOCOMP	7305.910	8.111e+03	2964.360	9272.620	108606.420	0.637	48.031	5.413
LOG(CEOCOMP)	8.489	1.003	7.994	9.135	11.596	-0.453	7.256	-1.380
CEOEQUITY	0.016	4.843e-02	0.001	0.009	0.686	0	70.974	7.352
LOG(CEOEQUITY) [7]	-5.725	1.774	-6.858	-4.699	-0.377	-13.999	0.671	-0.061
CEOTENURE	8.000	7.617	2.499	11.225	55.083	0.071	4,438	1.830
LOG(CEOTENURE)	1.625	1.023	0.916	2.418	4.009	-2.643	-0.339	-0.328

<표3> SIZE 등 8개 변수의 로그변환 전후 기초통계량

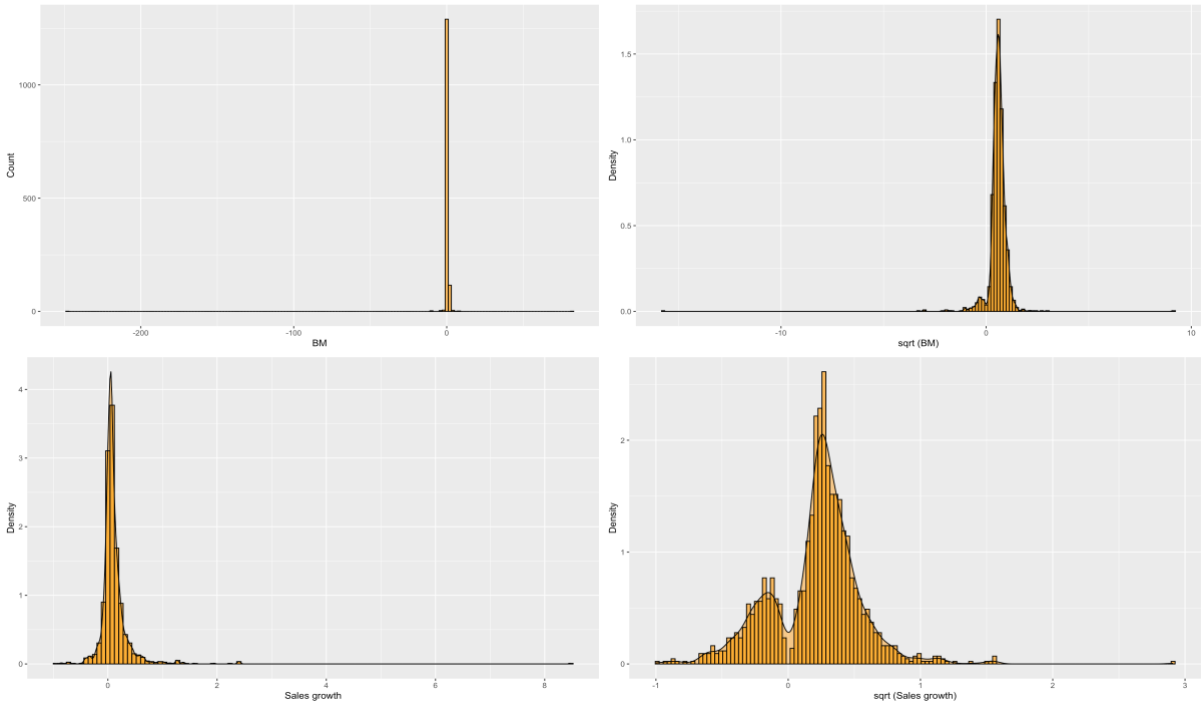
한편, BM과 SALESGROWTH 변수는 제곱근변환을 진행하였다. 이들 변수는 음수도 가지므로 로그변환을 할 수 없었다. 그러므로 해당 변수의 절대값에 제곱근을 취한 후, 원래의 부호를 재부여하는 방식으로 변환을 진행하였다. 다음의 <그림3>은 BM과 SALESGROWTH 변수의 제곱근 변환 전후 히스토그램을 나타내며, <표4>에서 변환 전후의 기초통계량을 파악할 수 있다.

[4] -Inf 값은 제외하였다.

[5] 4와 동일

[6] 4와 동일

[7] 4와 동일

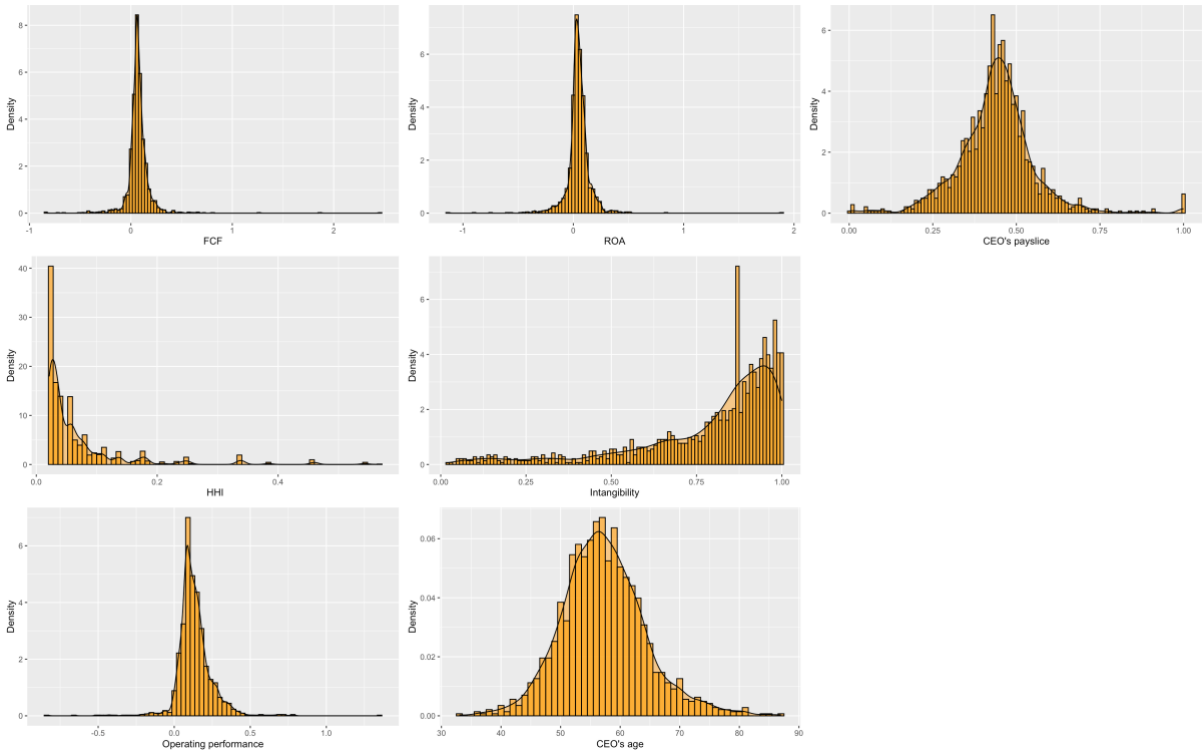


<그림3> BM(상), SALESGROWTH(하) 제곱근변환 전후 히스토그램

변수명	평균	표준편차	Q1	Q3	최댓값	최솟값	첨도	왜도
BM	0.299	6.976	0.189	0.585	82.741	-248.424	1141.245	-30.560
SQRT(BM)	0.569	0.643	0.435	0.765	9.096	-15.762	313.873	-10.471
SALESGROWTH	0.113	0.340	0.005	0.153	8.470	-1.000	260.664	11.890
SQRT(SALESG.)	0.213	0.339	0.068	0.391	2.910	-1.000	3.755	0.187

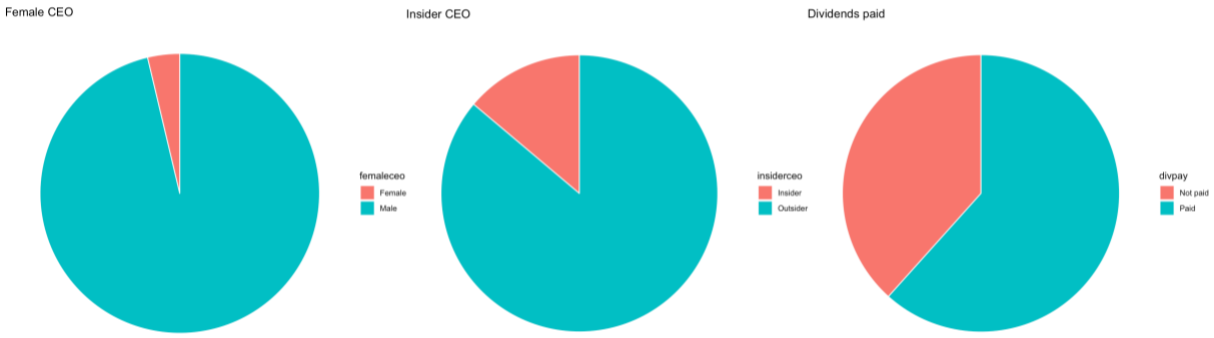
<표4> BM, SALESGROWTH 제곱근변환 전후 기초통계량

그 외 나머지 연속형 변수는 추가적인 변환이 필요 없거나, 로그변환 등의 변환을 통해서도 자료의 비대칭성을 해소하기 어려웠다. FCF, ROA, OPPERF, CEOPAYSLICE, 그리고 CEOAGE 변수의 경우 변환 없이도 분포의 비대칭성이 뚜렷하지 않았으며, HHI와 INTAN 변수의 경우 비대칭성을 줄일 수 있는 변환 방법을 찾기 어려웠다. 따라서 이들 변수는 변환없이 분석에 사용하고자 한다. 다음 <그림4>는 이 변수들의 히스토그램을 나타낸다.



<그림4> FCF 등 7개 연속형 변수에 대한 히스토그램

한편, 본 연구는 DIVPAY, FEMALECEO 등 범주형 변수도 분석의 대상으로 포함한다. 따라서 이들에 대한 통계량을 살펴보는 것도 중요한데, <그림5>는 범주형 변수인 FEMALECEO, INSIDERCEO, 그리고 DIVPAY의 비율을 나타낸 파이도표(Pie-chart)이다. 해당 도표를 통해 각 변수별 항목의 비율을 시각적으로 파악할 수 있다. 범주형 변수의 빈도수와 비율에 대한 정확한 정보는 <표6>에 기술하였다.



<그림5> 범주형 변수의 파이도표

다음의 <표5>는 본 연구의 분석에 사용하는 연속형 변수의 기초통계량을 요약한 것이다. <표6>는 범주형 변수의 범주별 빈도와 비율을 정리한 표이다. 변수의 성격에 따라 기업의 재무적 특성을 나타내는 변수와 CEO 특성을 나타내는 변수를 분류하여 기술하였다. 한편 기업 식별을 위한 변수 중 STATE는 범주가 상당히 많아 상관분석을 진행하는 과정에서 자세히 기술하도록 하였다.

변수명	평균	표준편차	Q1	Q3	최댓값	최솟값	첨도	왜도
<기업의 재무적 특성을 나타내는 변수>								
LOG(SIZE)	8.098	1.745	6.941	9.163	15.023	2.483	0.470	0.333
LOG(LEVERAGE)[8]	-1.951	1.266	-2.319	-1.217	0.435	-9.648	8.524	-2.336
LOG(TOBINSQ)	0.613	0.544	0.212	0.914	2.973	-1.281	1.105	0.869
LOG(CASHRATIO)[9]	-2.568	1.421	-3.418	-1.543	0.458	-11.645	3.454	-1.068
LOG(INVEST) [10]	-1.649	0.798	-2.129	-1.171	3.530	-6.430	3.174	-0.060
SQRT(BM)	0.569	0.643	0.435	0.765	9.096	-15.762	313.873	-10.471
SQRT(SALESG.)	0.213	0.339	0.068	0.391	2.910	-1.000	3.755	0.187
FCF	0.073	0.140	0.329	0.106	2.447	-0.844	85.110	4.820
HHI	0.068	0.077	0.026	0.073	0.566	0.021	12.951	3.296
OPPERF	0.133	0.120	0.080	0.178	1.363	-0.832	16.421	0.656
ROA	0.045	0.120	0.010	0.087	1.878	-1.128	50.972	1.169
INTAN	0.796	0.212	0.722	0.946	1.000	0.015	2.230	-1.631
RNDRATIO	0.039	0.069	0	0.050	0.809	0	25.756	4.033
<CEO 특성을 나타내는 변수>								
CEOAGE	33.000	7.060	53.000	61.000	87.000	33.000	0.996	0.407
LOG(CEOCOMP)	8.489	1.003	7.994	9.135	11.596	-0.453	7.256	-1.380
LOG(CEOEQUITY)[11]	-5.725	1.774	-6.858	-4.699	-0.377	-13.999	0.671	-0.061
LOG(CEOTENURE)	1.625	1.023	0.916	2.418	4.009	-2.643	-0.339	-0.328
CEOPAYSICE	0.440	1.160e-01	0.381	0.494	1.000	0.001	4.312	0.520

<표5> 연속형 변수의 기초통계량

[8] 4와 동일

[9] 4와 동일

[10] 4와 동일

[11] 4와 동일

변수명	내용	빈도수	비율	변수명	내용	빈도수	비율
<기업의 재무적 특성을 나타내는 변수>							
DIVPAY	배당금 지급	881	0.617				
	배당금 미지급	548	0.383				
<CEO 특성을 나타내는 변수>							
INSIDERCEO	내부승진	198	0.139	FEMALECEO	여	53	0.037
	외부채용	1231	0.861		남	1376	0.963

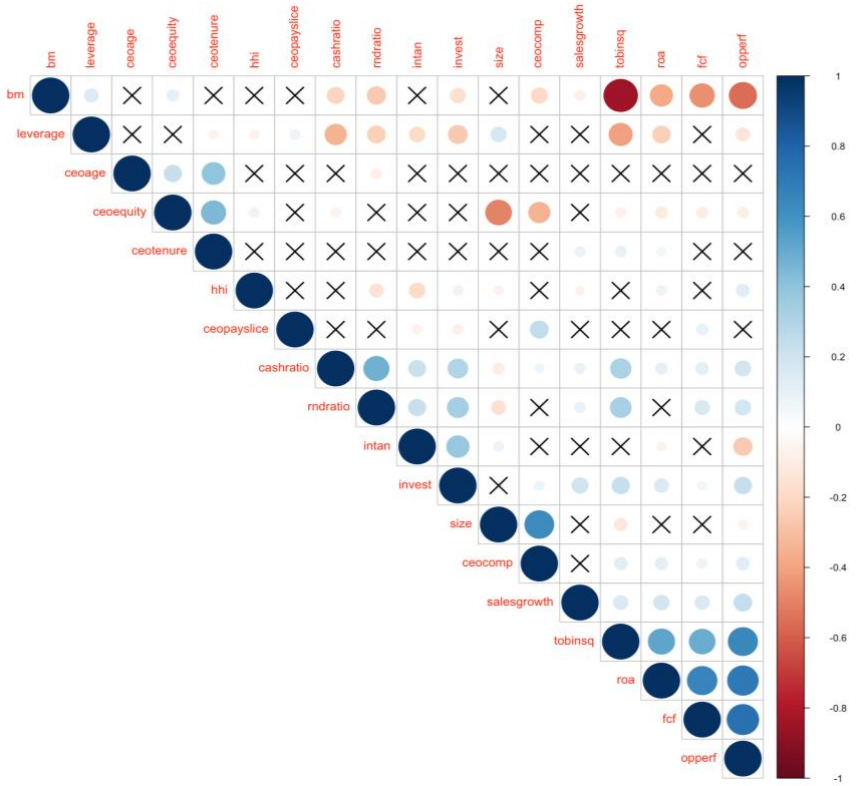
<표6> 범주형 변수의 빈도수, 비율

나. 종속변수와 설명변수간 상관분석

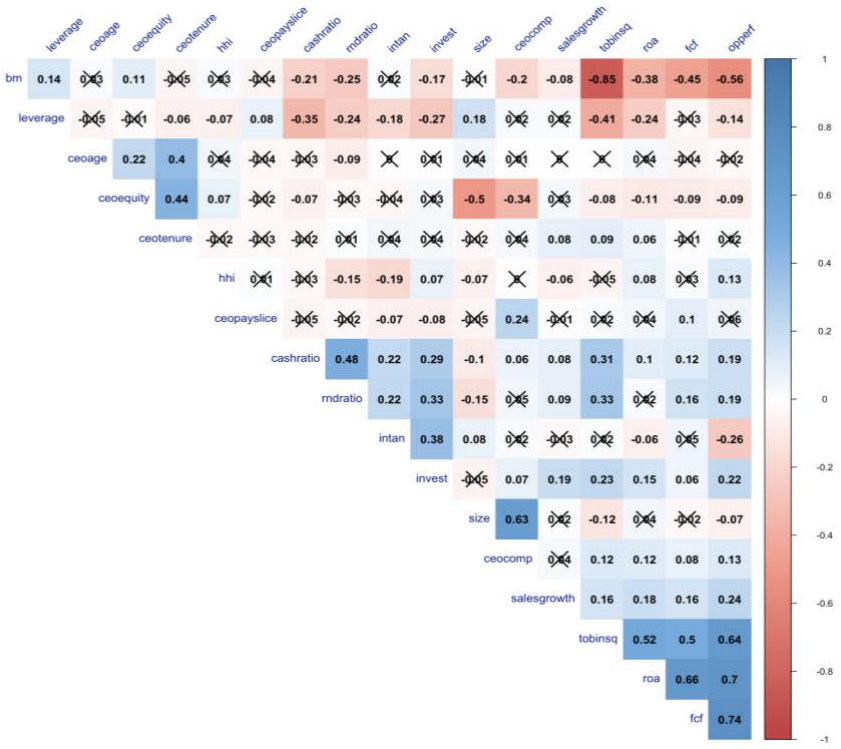
종속변수인 RNDRATIO와 모든 설명변수간 상관분석을 진행하였다. RNDRATIO는 연속형 변수이므로 설명변수의 유형에 따라 상이한 방법으로 변수간 상관성을 분석하였다. 먼저, 설명변수도 연속형 변수인 경우 상관계수에 대한 가설검정을 진행하였다. 종속변수가 정규분포를 따르지 않으므로 Pearson이 아닌 Kendall 상관계수와 Spearman 상관계수를 계산하였으며, 상관관계가 없다는 귀무가설을 설정하고 검정(cor.test)을 진행하였다. ROA, LOG(CEOCOMP), LOG(CEOEQUITY), LOG(CEOTENURE), 그리고 CEOPAYSLICE 변수는 두 상관계수 중 어느 것을 기준으로도 RNDRATIO와 상관성을 찾을 수 없었으며, 따라서 분석에서 제외하였다. 이들 변수의 상관계수와 검정으로부터의 P값은 <표7>에 정리하였다. 또한 <그림6-1>과 <그림6-2>는 연속형 변수 사이의 상관성을 요약한 행렬이다. <그림6-1> 행렬 내부의 원의 크기와 색에 따라서 상관성을 시각적으로 파악할 수 있는데, 원이 클수록 변수 간 더 높은 상관성을 가진다는 의미이다. 원의 색깔이 파란색에 가까울수록 양의 상관관계를 보이며, 빨간색에 가까울수록 음의 상관관계를 보이게 된다. <그림6-2>에서는 Spearman 상관계수를 행렬 내부에 기술하여 두 변수 간 상관성을 수치적으로 파악할 수 있도록 하였다. 행렬 내부의 색깔이 파란색에 가까울수록 두 변수 간 양의 상관성을 나타내며, 빨간색에 가까울수록 음의 상관성을 의미한다. Spearman 상관계수를 기준으로 상관성이 없는 변수는 Kendall 상관계수를 기준으로 제거된 <표7>의 변수와 동일했다.

변수명	Kendall 상관계수	상관계수 검정 P값
<기업의 재무적 특성을 나타내는 변수>		
ROA	0.025	0.174
<CEO 특성을 나타내는 변수>		
LOG(CEOCOMP)	-0.004	0.842
LOG(CEOEQUITY)	0.012	0.493
LOG(CEOTENURE)	0.028	0.122
CEOPAYSLICE	-0.027	0.144

<표7> 종속변수와 상관성이 없는 연속형 설명변수

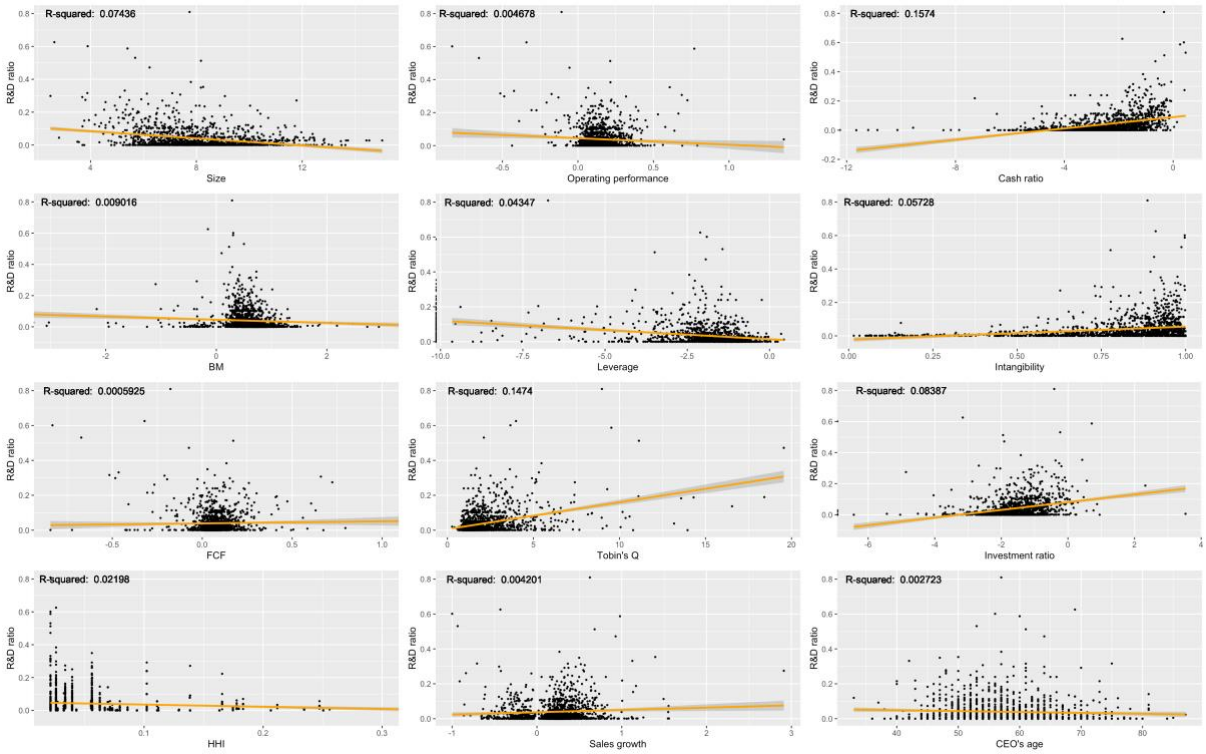


<그림6-1> 연속형 변수간 Correlation matrix



<그림6-2> 연속형 변수간 Correlation matrix with 상관계수

한편 <그림7>은 상관분석에서 종속변수와 상관성을 가지는 독립변수의 단순회귀분석 시행 후 산점도를 나타낸 것이다. 그림에는 R-squared값을 표시하였는데, 단일 독립변수가 종속변수에 대해 큰 설명력을 가지는 사례는 찾을 수 없었다. <표8>에서는 종속변수와 상관성을 가지는 독립변수의 단순회귀분석 결과를 요약하였다. 한편 변환과정을 거친 변수에 대해서도 편의상 변수의 기존 이름을 그대로 사용하였다. 예를 들어 SIZE 변수의 경우 로그변환을 통해 비대칭성을 줄였지만, <그림6-1>, <그림6-2>, 그리고 <그림7>에서는 LOG(SIZE)가 아닌 SIZE로 표기하였다. 뒤이어 나오는 <그림>에서도 편의상 변수의 기존 이름을 그대로 사용하기로 한다.

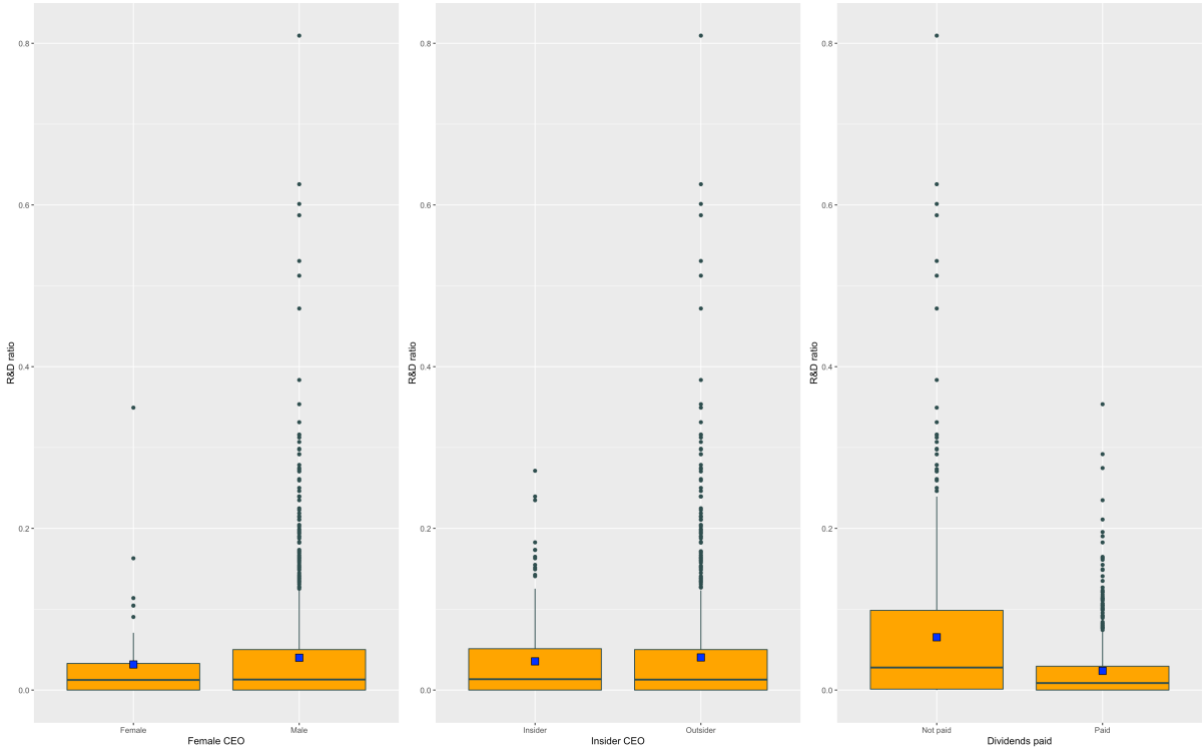


<그림7> 연속형 설명변수와 종속변수간 산점도

변수명	계수	P값	R-squared
<기업의 재무적 특성을 나타내는 변수>			
LOG(SIZE)	-0.011	2e-16	0.074
SQRT(BM)	-0.010	0.0003	0.009
FCF	0.012	0.358	0.0006
HHI	-0.134	1.8e-08	0.022
OPPERF	-0.039	0.009	0.005
LOG(LEVERAGE)	-0.011	3.04e-14	0.043
LOG(TOBINSQ)	0.015	2e-16	0.147
SQRT(SALESGROWTH)	0.013	0.014	0.004
LOG(CASHRATIO)	0.019	2e-16	0.157
INTAN	0.078	2e-16	0.057
LOG(INVEST)	0.025	2e-16	0.084
<CEO 특성을 나타내는 변수>			
CEOAGE	-0.0005	0.049	0.003

<표8> 연속형 설명변수와 종속변수간 단순회귀분석 결과

다음으로 종속변수 RNDRATIO와 범주형 변수간 상관분석을 진행하였다. 본 연구의 자료는 STATE, DIVPAY, INSIDERCEO, 그리고 FEMALECEO 4가지의 범주형 변수를 포함한다. DIVPAY, INSIDERCEO, 그리고 FEMALECEO의 경우 Boxplot을 그려 두 변수 사이의 상관성을 시각적으로 확인하였다. 이는 <그림8>로 제시하였다. 이후 변수 내 집단의 분산이 같다는 귀무가설을 F검정을 이용하여 검정하였으며, 이 결과를 바탕으로 평균의 차이에 대해 T검정(귀무가설: 평균이 같다)를 실시하였다. 우선 DIVPAY와 INSIDERCEO의 경우 F검정 결과 각각 P값이 2.2e-16과 3.99e-09이며, 따라서 변수 내 두 집단의 분산이 서로 다르다는 결론을 내릴 수 있었다. 이후 평균의 차이에 대하여 Welch Two Sample t-test를 진행하였다. DIVPAY의 P값은 0.303, INSIDERCEO의 P값은 0.244로 두 변수 내 집단의 평균이 같다는 결론을 내릴 수 있었다. 따라서 DIVPAY와 INSIDERCEO는 변수 내 집단에 따른 RNDRATIO와의 상관성이 없는 것으로 판단되어 분석에서 제외하였다. 한편 FEMALECEO의 경우 F검정 결과 P값이 0.064을 가져 분산이 서로 같다고 판단했다. 이를 통해 Two Sample t-test를 진행하였는데, P값이 0.389를 나타내 CEO의 성별에 따른 RNDRATIO 평균이 같다고 결론지을 수 있었다. 따라서 FEMALECEO 변수도 RNDRATIO와 상관성이 없는 것으로 판단하여 분석에서 제외하였다. 위의 과정을 <표9>에 간단히 정리하였다.

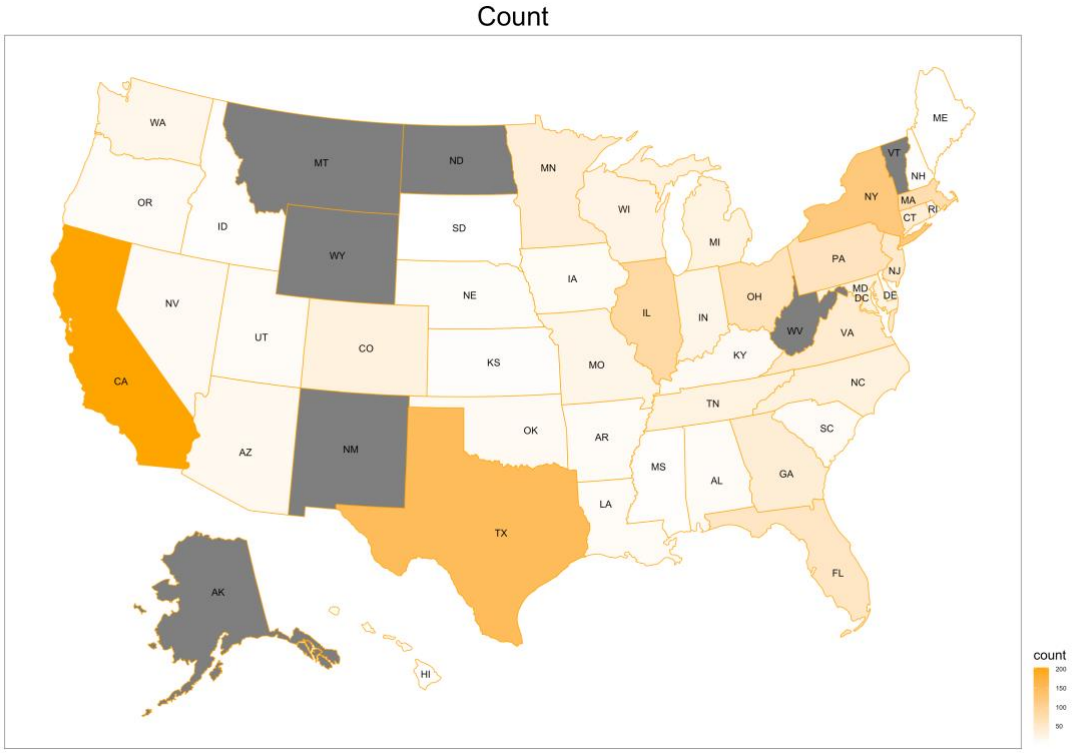


<그림8> 범주형 설명변수와 종속변수 간 Boxplot

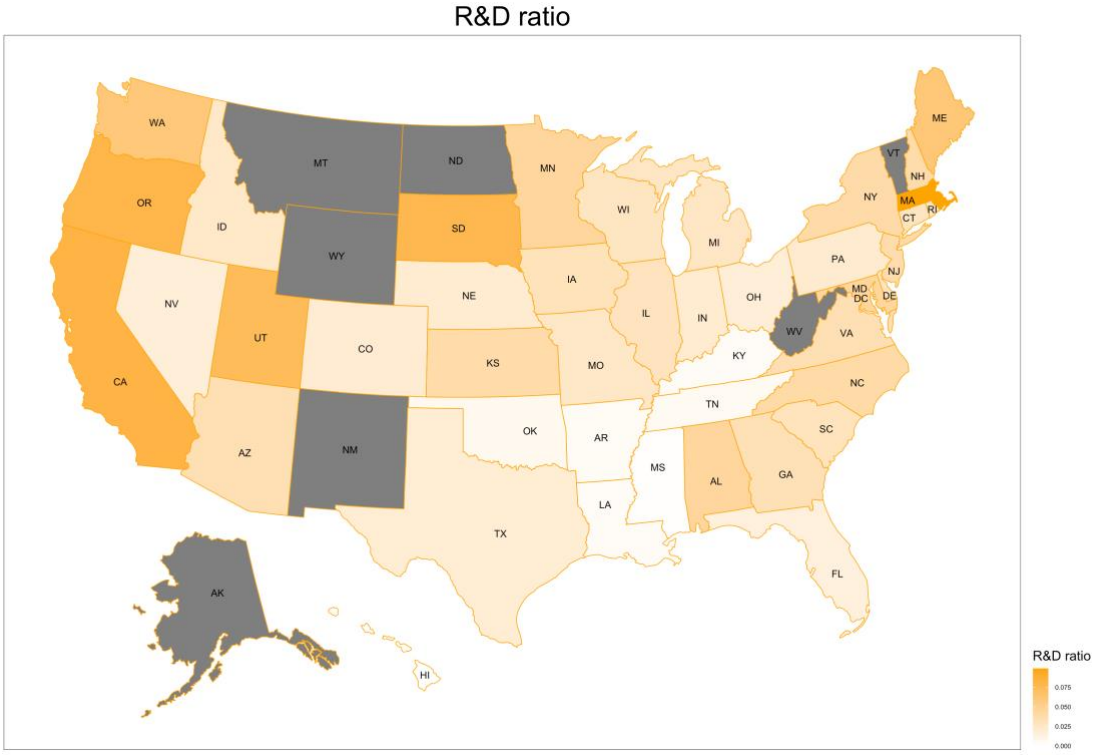
변수명	F검정 P값	T검정 P값	
<기업의 재무적 특성을 나타내는 변수>			
DIVPAY	2.2e-16	0.303	이분산, 등평균
<CEO 특성을 나타내는 변수>			
INSIDERCEO	3.99e-09	0.244	이분산, 등평균
FEMALECEO	0.06	0.389	등분산, 등평균

<표9> 범주형 설명변수 상관분석

마지막으로 범주형 설명변수 STATE와 종속변수 간 상관성을 분석하였다. STATE 변수는 51개 항목으로 분류가 되어있었다. STATE와 RNDRATIO를 시각적으로 살펴보았을 때, 기업 수가 상대적으로 많은 CA(California)와 MA(Massachusetts)의 경우 RNDRATIO 항목도 높게 나오는 것으로 나타났다. 따라서 우선적으로 STATE를 크게 CA, MA, 그리고 그 외($(CA \cup MA)^C$)의 항목으로 나누어 분석을 진행하였다. 이는 <그림9-1>과 <그림9-2>를 통해 직관적으로 파악할 수 있다.

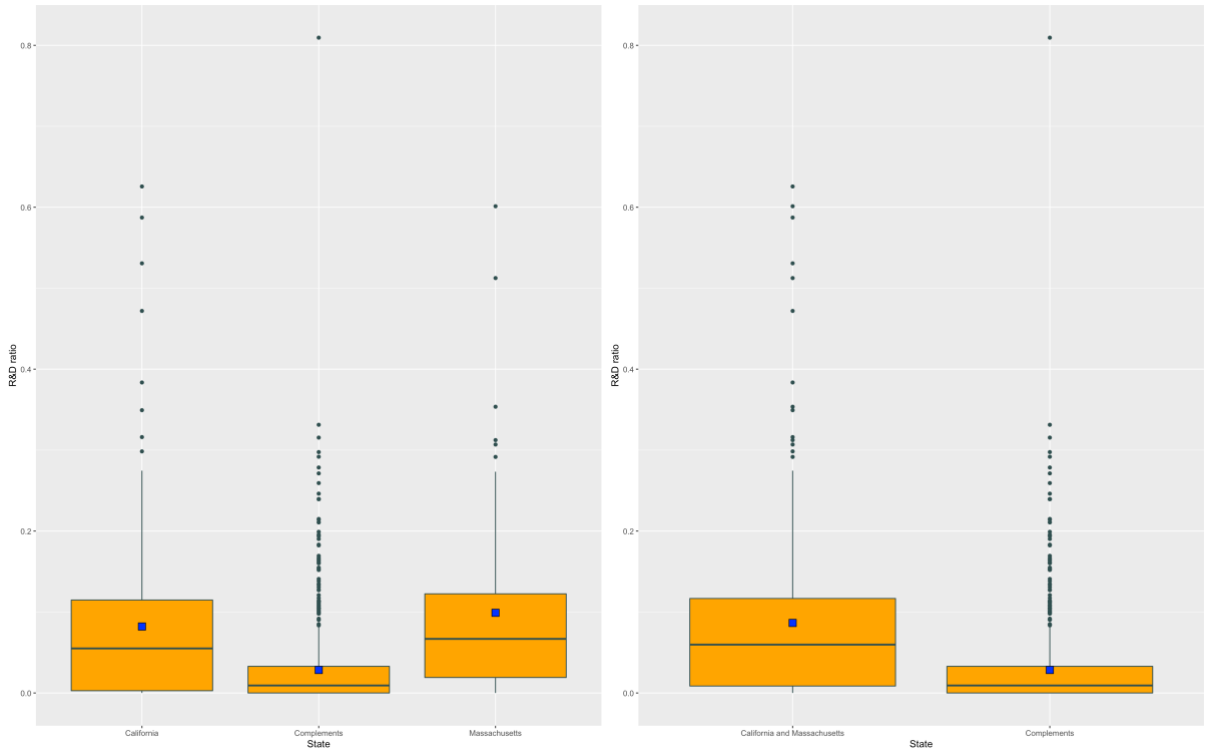


<그림9-1> STATE별 기업의 수



<그림9-2> STATE별 평균 R&D 투자 비율

다음으로 세 가지 항목으로 나눈 STATE에 대한 Boxplot을 그림으로써 STATE 변수 내 집단별로 RNDRATIO에 대한 분산과 평균의 차이가 존재하는지 살펴보고자 하였다. F검정과 T검정 결과 CA와 MA는 분산과 평균이 모두 같았으며, 따라서 이질적인 집단으로 분류하기 어려웠다. 그러나 CA와 $(CA \cup MA)^C$), 그리고 MA와 $(CA \cup MA)^C$의 경우는 검정 결과 분산과 평균이 모두 다른 것으로 나타났다. 따라서 우리는 세 가지 집단으로 대분류를 하는 대신, CA와 MA를 통합하여 STATE 변수를 두 가지 집단으로 재분류하기로 하였다. 한편, STATE 변수 내 $(CA \cup MA)$ 와 $(CA \cup MA)^C$ 집단의 경우 검정 결과 분산과 평균이 모두 다른 것으로 결론 내릴 수 있었다. 따라서 우리는 STATE를 두 가지 항목으로 분류하여 분석에 포함하였다. 이들 집단에 대한 Boxplot은 <그림10>에서 볼 수 있다.



<그림10> STATE를 세 가지 항목으로 분류한 경우(좌)와 두 가지로 분류한 경우(우)의 Boxplot

종속변수와 설명변수간 상관분석을 통해 결과적으로 다음의 변수를 종속변수와 상관성이 있는 것으로 정리할 수 있었다. 종속변수인 RNDRATIO를 제외하고 13개 변수이며, 범주형 변수 1개와 연속형 변수 12개로 구성된다.

변수명	변수의 성격
<기업 식별을 위한 변수>	
STATE	범주형
<기업의 재무적 특성을 나타내는 변수>	
LOG(SIZE)	연속형
SQRT(BM)	연속형
FCF	연속형
HHI	연속형
OPPERF	연속형
LOG(LEVERAGE)	연속형
LOG(TOBINSQ)	연속형
SQRT(SALESGROWTH)	연속형
LOG(CASHRATIO)	연속형
INTAN	연속형

LOG(INVEST)	연속형
RNDRATIO	연속형 / 종속변수
<CEO 특성을 나타내는 변수>	
CEOAGE	연속형

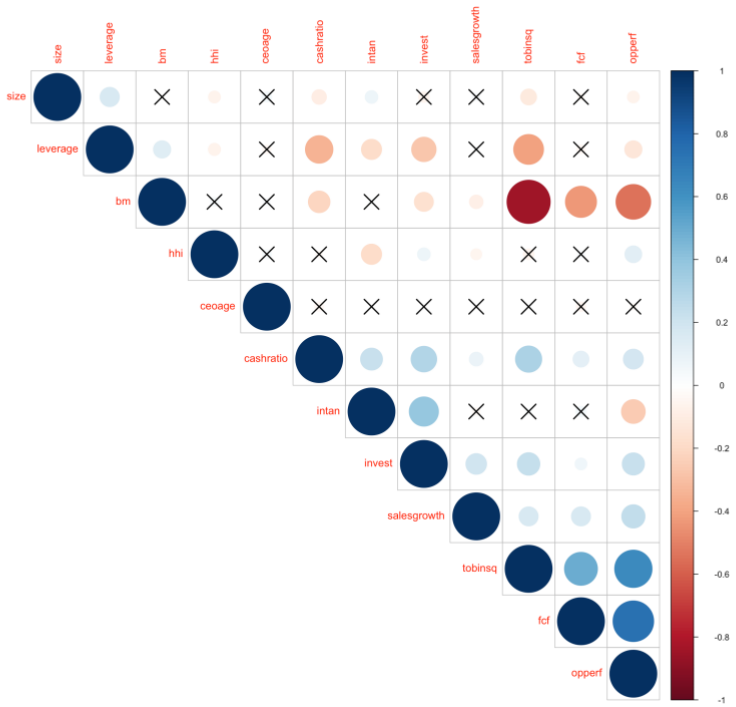
<표10> 종속변수와 상관성이 있는 변수

다. 설명변수간 상관분석

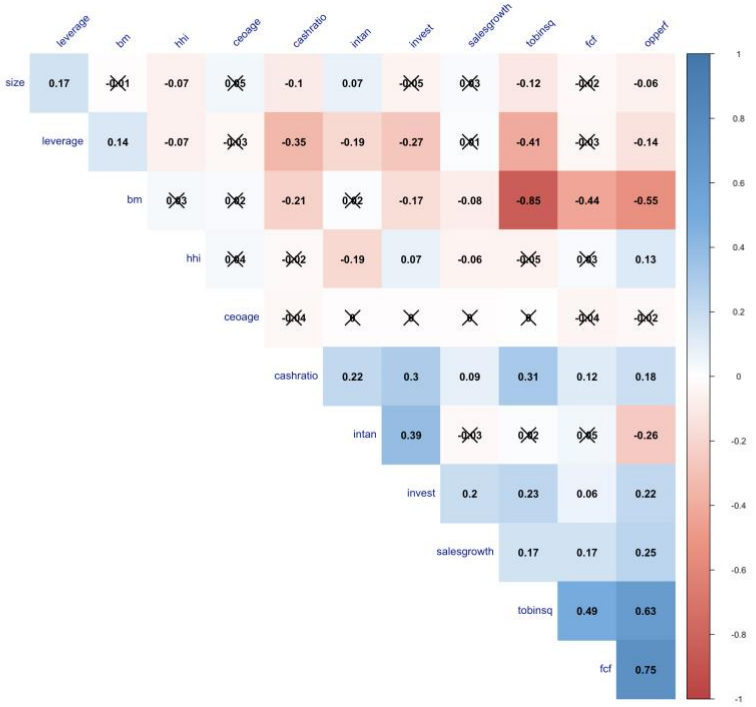
회귀분석 시 설명변수간 독립성을 확보하는 것은 중요하다. 이 가정은 회귀분석을 통해 구해지는 최소제곱해(Least Square Solution)의 유일성을 보장하기 위하여 필수적이며, 가정이 어긋나는 경우 다중공선성(Collinearity) 문제를 일으킨다.[12] 다중공선성이 존재하는 경우 회귀모형이 F검정을 통해서는 유의하게 나타나지만, 모든 변수는 T검정 상 유의성을 확보하지 못한다. 또한 회귀분석을 통해 추정한 계수의 부호나 값이 선행연구나 탐색적 자료분석에서 추정한 값과 다르게 나타난다. 따라서 자료분석을 하는 과정에서 독립변수간 상관성을 분석하는 작업도 매우 중요하며, 본 연구는 앞서 정리한 13개의 변수에 대해서 상관분석을 진행하고자 한다.

우선, 연속형 변수간 Correlation matrix를 그릴 수 있었다. <그림11-1>에서는 행렬 내부의 원의 크기와 색에 따라서 상관성을 시각적으로 파악할 수 있는데, 원이 클수록 변수 간 더 높은 상관성을 가진다는 의미이다. 원의 색깔이 파란색에 가까울수록 양의 상관관계를 보이며, 빨간색에 가까울수록 음의 상관관계를 보이게 된다. 한편 <그림11-2>에서는 Spearman 상관계수를 행렬 내부에 기술하여 두 변수간 상관성을 수치적으로 파악할 수 있도록 하였다. 행렬 내부의 색깔이 파란색에 가까울수록 두 변수간 양의 상관관계를 나타내며, 빨간색에 가까울수록 음의 상관관계를 의미한다. 한편 <그림11-1>과 <그림11-2>의 두 행렬에서 동일한 변수집단에 대해 'X'표시가 나타난 것을 볼 수 있다. 이는 상관성에 대한 가설검정 결과, 상관관계가 없다는 귀무가설을 기각하지 못하여 두 변수간 상관성이 없다고 결론 내릴 수 있는 집단을 의미한다. 특히 CEOAGE의 경우 다른 설명변수와 상관성이 전혀 없는 것으로 나타났다. 반면, OPPERF의 경우 기업의 재무적 특성을 나타내는 다른 설명변수와 어느정도 상관성이 있음을 파악할 수 있다.

[12] Chatterjee, S. and Hadi, A.S. (2012) Regression Analysis by Example. 5th Edition, Wiley, New York, 98.

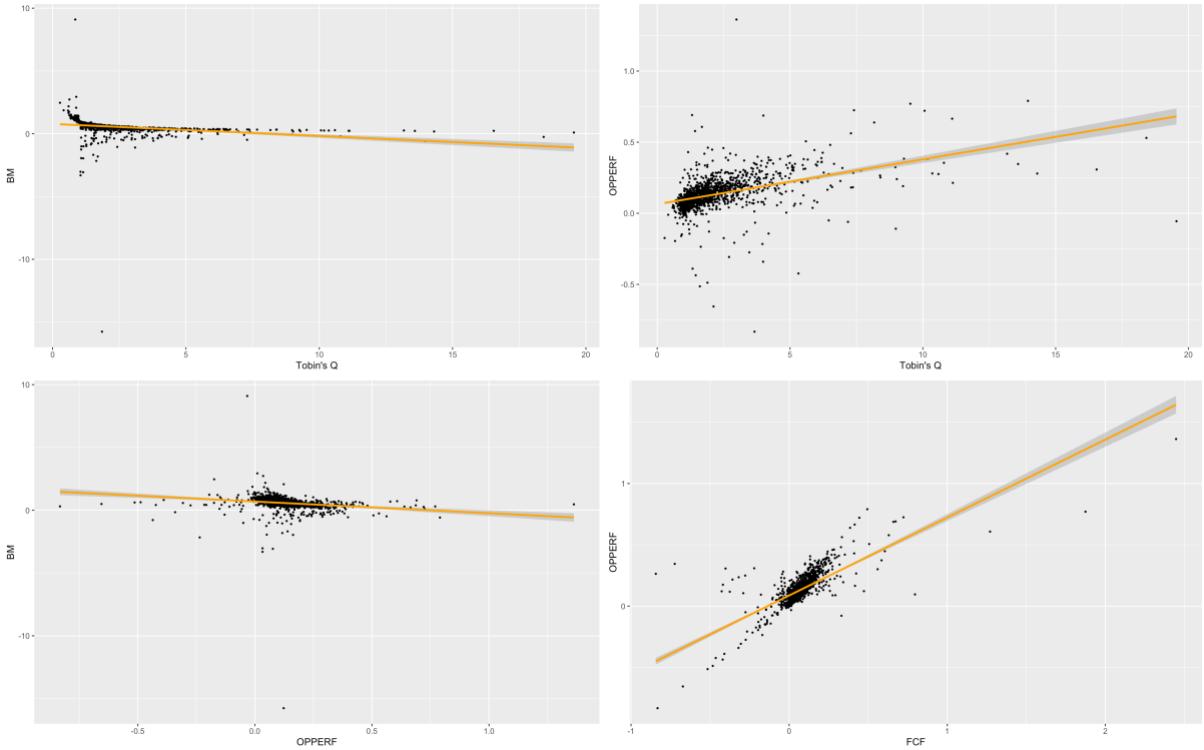


<그림11-1> 연속형 설명변수간 Correlation matrix



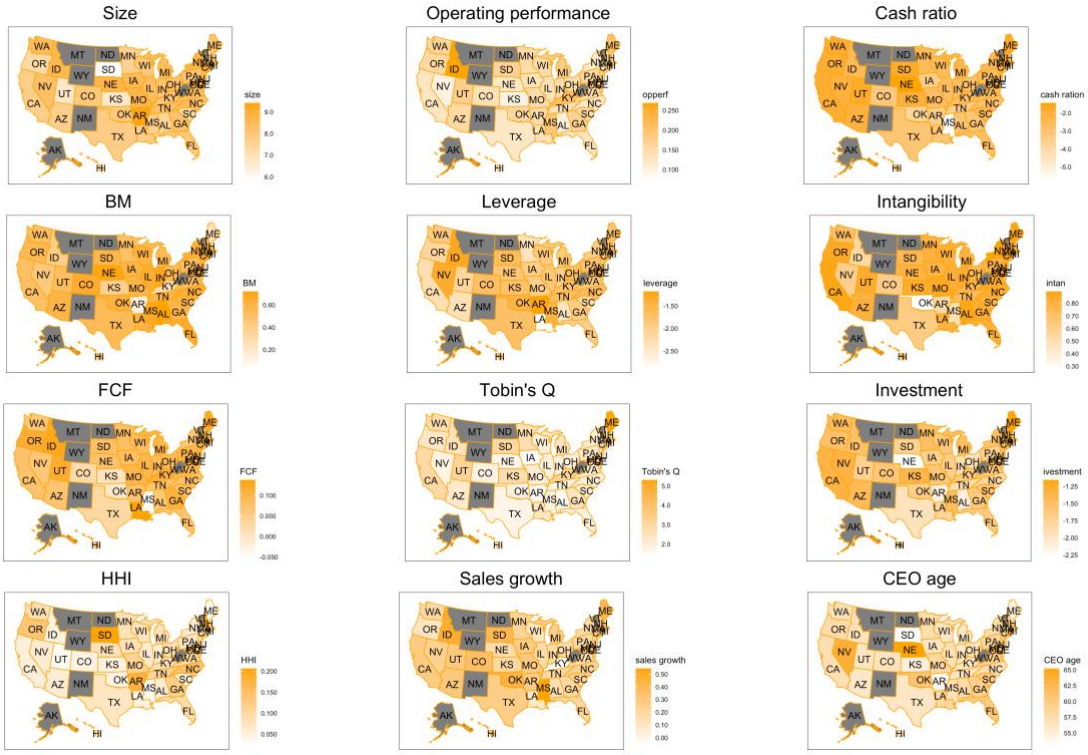
<그림11-2> 연속형 설명변수간 Correlation matrix with 상관계수

본 연구는 상관계수의 절대값이 0.5보다 높은 변수집단에 대해 산점도를 그려보았다. 산점도를 통해 두 변수 사이에 상관성이 존재하는지 좀 더 직관적으로 파악할 수 있다. <그림12>를 보면 OPPERF, BM, 그리고 TOBINSQ 변수의 경우 서로 상관성을 가질 가능성이 있으며, OPPERF와 FCF의 경우도 상관성이 존재할 여지가 있다. 따라서 추후 다중회귀분석에 포함할 변수를 선택하는 과정에서 다중공선성에 대한 검사를 필수적으로 진행해야 한다. 예를 들어 모든 설명변수의 VIF(Variance Inflation Factor)를 계산하여 VIF 값이 10을 넘으면서 가장 큰 변수를 순차적으로 하나씩 제외하면서 다중공선성 문제를 해결할 수 있다.



<그림12> 상관계수 절대값이 0.5보다 큰 설명변수간 산점도

한편, 범주형 변수인 STATE와 연속형 설명변수간 상관성을 분석하고자 하였다. STATE는 $(CA \cup MA)$와 $(CA \cup MA)^C$ 두 집단으로 분류하였는데, 만약 STATE와 다른 설명변수 사이 상관성이 있다면 CA와 MA 지역에서만 상대적으로 높은 수치를 보이거나, 반대로 낮은 수치를 보일 것이다. 그러나 <그림13>을 살펴보았을 때 어느 설명변수에 대해서도 위와 같은 경향성을 찾기 어려웠다. 따라서 STATE와 다른 연속형 설명변수간 두드러지는 상관성은 없는 것으로 파악하였다.



<그림13> STATE와 연속형 설명변수의 상관성

제 3장 결론

제 1절 분석 결과

본론에서 각 변수의 기초통계량을 바탕으로 변수의 비대칭성을 정리할 수 있었으며, 종속변수와 설명변수간 상관분석과 설명변수간 상관분석을 진행하였다. SIZE, LEVERAGE, TOBINSQ, CASHRATIO, INVEST, CEOCOMP, CEOEQUITY, 그리고 CEOTENURE 변수는 로그변환을 통해 비대칭성을 줄일 수 있었다. 한편 BM과 SALESGROWTH 변수는 제곱근변환을 통해 분석에 용이하도록 변수를 가공하였다. FCF, ROA, OPPERF, CEOPAYSLICE, 그리고 CEOAGE 변수의 경우 원 자료의 분포에서 비대칭성을 찾기 어려웠으며, HHI와 INTAN 변수는 비대칭적 분포를 보였으나 변환을 통해 이를 해결할 수 없었다. 따라서 이들 변수는 변환없이 분석에 사용하였다. 또한 종속변수인 RNDRATIO도 정규성을 가진다고 가정하기 어려워, 본 연구에서는 로그변환, 제곱근변환, 그리고 ARCSINE변환을 고려하였다. 그러나 변환을 통해서도 정규성을 확보할 수 없었으며, RNDRATIO가 0인 집단을 분석에서 제외하는 것도 불가능하여 변환없이 분석을 진행하였다.

한편, 종속변수와 설명변수간 상관분석을 통해 상관성이 없는 설명변수를 분석에서 제외하였다. 먼저 연속형 설명변수의 경우 상관계수에 관한 검정을 진행하였는데, ROA, CEOCOMP, CEOEQUITY, CEOTENURE, 그리고 CEOPAYSLICE를 분석에서 제외할 수 있었다. 한편 범주형 설명변수는 분산의 차이에 대한 F검정과 평균이 차이에 대한 T검정을 실시하였고, DIVPAY, INSIDERCEO, 그리고 FEMALECEO 변수를 분석에서 제외하였다. 범주형 변수 STATE는 변수 내 CA와 MA 집단에서 종속변수 RNDRATIO와 상관성을 보여 일차적으로 세 집단으로 분류하였다. 이후 이들간 F검정과 T검정을 통해 $(CA \cup MA)$와 $(CA \cup MA)^C$ 두 집단으로 분류할 수 있었다.

마지막으로 설명변수간 상관분석을 진행하였다. 일차적으로 연속형 설명변수간 Correlation matrix를 통해 시각적으로 상관성을 파악하였다. CEOAGE의 경우 다른 설명변수와 상관성이 없는 것으로 파악되었지만, OPPERF의 경우 기업의 재무적 특성을 나타내는 다른 설명변수와 어느정도 상관성을 보이는 것으로 나타났다. 한편 Spearman 상관계수가 큰 변수집단에 대해서는 추가적으로 산점도를 그렸는데, OPPERF, BM, 그리고 TOBINSQ가 서로 상관성이 있을 가능성이 있었다. 이를 통해 이후 다중회귀분석을 진행하기에 앞서 변수를 선택하는 과정에서 다중공선성에 대한 검사를 필수적으로 진행해야 함을 알 수 있었다. 예를 들어 각 설명변수의 VIF를 조사하여 해당 수치가 10 이상의 값을 가지는 동시에 가장 큰 변수를 분석에서 순차적으로 제외함으로써 이를 해결할 수 있다. 마지막으로 범주형 설명변수 STATE의 경우 다른 설명변수와 상관성이 없는 것으로 파악할 수 있었다.

제 2절 향후 분석방향 제시

제 1절에서 다루고자 했던 세 가지 질문을 본 연구를 통해 답할 수 있었다. INVEST, TOBINSQ 등 기업의 재무적 특성을 나타내는 많은 변수가 기업의 R&D 투자와 상관성이 있는 것으로 나타났다. CEO의 특성을 나타내는 변수도 R&D 투자에 설명력을 가졌는데, CEO의 나이를 의미하는 CEOAGE 변수가 R&D 투자와 상관성이 있다고 파악할 수 있었다. 이후 다중회귀분석을 통해 CEOAGE와 R&D 투자의 상관성이 어느 정도인지 이해할 수 있을 것으로 기대한다. 두 변수가 어떤 형태의 상관성을 가지는지는 RNDRATIO를 종속변수로 하며 CEOAGE를 설명변수로 하는 다중회귀식을 통하여 파악할 수 있을 것이다. 다만 종속변수가 비율을 의미하는 만큼 CEOAGE의 계수를 해석하는 과정에서 주의가 필요하다. 한편 종속변수 RNDRATIO는 변수변환을 통해서도 정규성을 확보하기 어려웠는데, RNDRATIO 변수의 평균과 표준편차를 이용하여 변수를 정규화하는 방법을 고민해볼 수 있다. 위 방법을 통해 종속변수의 분포가 정규분포와 유사하게 된다면 이를 종속변수로 설정한 회귀분석을 고려해야 할 것이다.

본 연구를 통해 정리한 변수를 모두 다중회귀분석에 포함하는 대신 다중공선성에 대한 검사가 이루어져야 한다. 이는 설명변수간 상관분석을 진행하는 과정에서 상관성이 존재할 가능성을 포착하였기 때문이다. 다중공선성 문제는 모든 설명변수의 VIF를 계산한 후 VIF 값이 10 이상인 변수 중 VIF가 가장 큰 변수를 분석에서 순차적으로 하나씩 제외하면서 해결할 수 있다. 이를 통해 설명변수간 독립성을 확보하여 올바른 분석이 이루어질 수 있도록 해야 한다. 또한 분석을 진행하면서 변수선택과정을 거쳐 유의한 변수를 최종 다중회귀분석식에 포함해야 한다.

다중회귀분석을 진행한 후에도 잔차분석을 통해 회귀분석의 기본가정이 어긋나지 않는지 살펴보아야 한다. 만약 가정이 지켜지지 않는다면 변수변환 등 추가적인 조치를 통해 이를 해결해야한다. 또한 극단치(Outlier)와 영향치(Influence)를 적절히 제거함으로써 회귀분석이 올바르게 진행될 수 있도록 노력해야한다.

정리하자면 본 연구는 다음과 같은 분석방향을 제시하고자 한다.

1. VIF를 계산한 후 VIF 값이 10 이상인 변수 중 가장 높은 값을 가지는 변수를 하나씩 순차적으로 제거함으로써 다중공선성을 해결한다.

2. 1을 거쳐 정리된 변수를 대상으로 변수선택을 실시한다. 이는 통계적으로 가장 유의한 변수부터 하나씩 식을 포함하는 Forward Selection Procedure, 모든 변수를 포함한 후 가장 유의하지 않은 변수를 하나씩 제외하는 Backward Selection Procedure, 그리고 이 둘을 통합한 Stepwise Method 등으로 실현할 수 있다.

3. 다중회귀분석을 진행한 후 잔차분석을 실시한다. 회귀분석의 기본가정이 어긋나지 않는지

살펴본 후, 극단치와 영향치를 적절히 제거함으로써 회귀분석이 올바르게 이루어질 수 있도록 노력한다.

참고문헌

Chatterjee, Benmelech, E., & Frydman, C. (2015). Military ceos. *Journal of Financial Economics*, *117*(1), 43-59.

Barker III, V. L., & Mueller, G. C. (2002). CEO characteristics and firm R&D spending. *Management Science*, *48*(6), 782-801.

Serfling, M. A. (2014). CEO age and the riskiness of corporate policies. *Journal of Corporate Finance*, *25*, 251-273.

S. and Hadi, A.S. (2012) Regression Analysis by Example. 5th Edition, Wiley, New York, 98.

Kronmal, R. (1993). Spurious Correlation and the Fallacy of the Ratio Standard Revisited. *Journal of the Royal Statistical Society. Series A (Statistics in Society), 156*(3), 379-392. doi:10.2307/2983064

Kang H. (2013). The prevention and handling of the missing data. *Korean journal of anesthesiology*, *64*(5), 402–406. https://doi.org/10.4097/kjae.2013.64.5.402