2016학년도 제1학기
Multivariate Data Analysis
팀프로젝트 보고서

Twins Analysis

제 출 자 : 7조
1302085 이응경
1302124 최희원

제 출 일 : 2016. 06. 08

목 차

1. 자료설명

텍사스 프레스 대학에서 조사한 쌍둥이 연구 데이터로서 각 쌍둥이간의 성적과 부모의 교육수준, 가족의 소득 수준에 대한 정보가 나와 있다. 각 정보들의 의미는 다음과 같고, 범주형 변수와 연속변수는 다음과 같이 이산변수로 나타나있다.

Pairnum :Twin pair number
Sex : 성별
Zygosity : 쌍둥이의 접합병식
Moed : Mother's educational level
Faed : Father's educational level
Finc : Family income level
English : NMSQT Subtest: English
Math : NMSQT Subtest: Mathematics
SocSci : NMSQT Subtest: Social Science
NatSci : NMSQT Subtest: Natural Science
Vocab : NMSQT Subtest: Vocabulary

```
value sexfmt 1='male' 2='female';
value zygfmt 1='identical' 2='fraternal';
value edfmt 1='<= 8th grade'
            2='part high school'
            3='high school grad'
            4='part college'
            5='college grad'
            6='graduate degree';
value incfmt 1='< $5000'
             2='$5000 to $7499'
             3='$7500 to $9999'
             4='$10000 to $14999'
             5='$15000 to $19999'
             6='$20000 to $24999'
             7='>= $25000'
```

같은 쌍의 쌍둥이 간의 구별을 위해서 'even' 변수를 추가하여 데이터를 수정하였다.

pairnum	sex	zygosity	moed	faed	faminc	english	math	socsci	natsci	vocab	even
1	2	1	3	4	2	14	13	17	18	14	FALSE
1	2	1	3	4	2	11	14	15	10	12	TRUE
2	2	1	1	1	1	20	20	16	16	13	FALSE
2	2	1	1	1	1	17	19	13	13	14	TRUE
3	2	1	1	1	1	11	8	15	16	12	FALSE
3	2	1	1	1	1	16	13	13	8	15	TRUE
4	1	2	3	2	4	9	19	7	10	6	FALSE
4	1	2	3	2	4	8	16	15	17	11	TRUE
5	1	2	5	4	3	15	23	23	21	21	FALSE
5	1	2	5	4	3	15	13	13	20	19	TRUE
6	2	2	4	2	3	20	17	16	12	12	FALSE
6	2	2	4	2	3	19	18	13	18	15	TRUE
7	1	1	4	6	3	26	20	27	23	28	FALSE
7	1	1	4	6	3	25	20	29	24	30	TRUE
8	1	1	4	6	5	28	31	31	29	30	FALSE
8	1	1	4	6	5	27	28	28	27	30	TRUE

2. 목적

1) 문과/이과 과목 간의 상관관계가 있는가.

2) 성적 상/중/하 그룹 별로 요인의 차이가 있는가.

3) 부모의 교육수준 상/하 그룹 별로 성적의 차이가 있는가.

4) 가정의 소득수준 상/하 그룹 별로 성적의 차이가 있는가.

5) 남/여 성별 그룹 별로 요인/성적의 차이가 있는가.

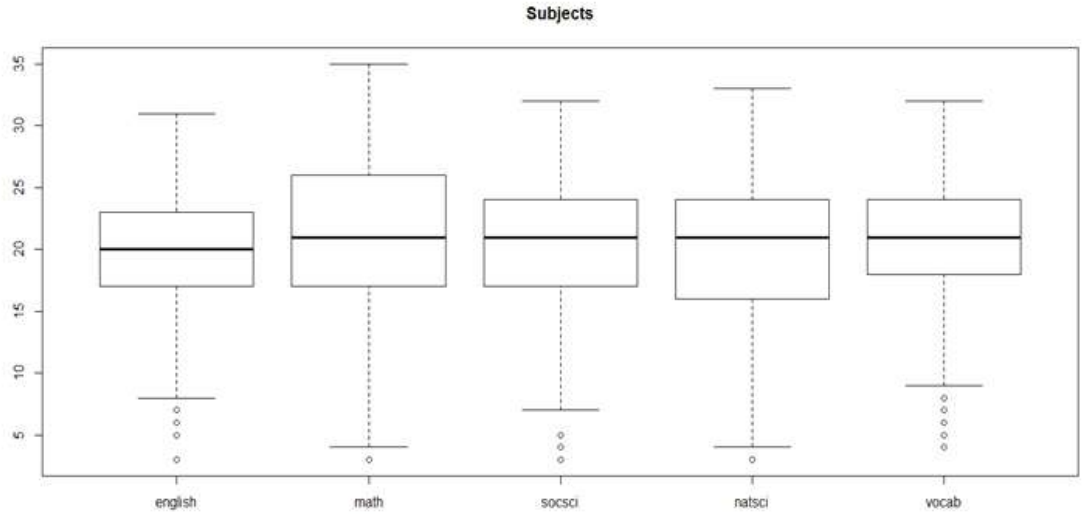
6) 쌍둥이 간의 성적에 차이가 있는가.

3. 분석

1) Summary statistics

```
> summary(twins[,4:11])
      moed            faed           faminc          english          math            socsci          natsci          vocab
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 3.00   Min.   : 3.00   Min.   : 3.00   Min.   : 3.00   Min.   : 4.00
 1st Qu.:3.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:17.00   1st Qu.:17.00   1st Qu.:17.00   1st Qu.:16.00   1st Qu.:18.00
 Median :3.000   Median :4.000   Median :3.000   Median :20.00   Median :21.00   Median :21.00   Median :21.00   Median :21.00
 Mean   :3.424   Mean   :3.587   Mean   :3.241   Mean   :19.76   Mean   :21.25   Mean   :20.69   Mean   :20.09   Mean   :21.06
 3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:23.00   3rd Qu.:26.00   3rd Qu.:24.00   3rd Qu.:24.00   3rd Qu.:24.00
 Max.   :6.000   Max.   :6.000   Max.   :7.000   Max.   :31.00   Max.   :35.00   Max.   :32.00   Max.   :33.00   Max.   :32.00
```

```
             moed      faed    faminc
moed    1.0000000 0.5582127 0.4010453
faed    0.5582127 1.0000000 0.5230169
faminc  0.4010453 0.5230169 1.0000000
```



각 요인의 분포는 위와 같다.

부모의 교육수준과 가정의 소득수준 간에 양의 상관관계가 있는 것을 알 수 있다.

2) Principal Components Analysis

```
> summary(princomp(subjects))
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5
Standard deviation     9.918813 4.0736321 3.35238377 2.93207343 2.22615435
Proportion of Variance 0.703898 0.1187282 0.08040772 0.06150919 0.03545691
Cumulative Proportion  0.703898 0.8226262 0.90303391 0.96454309 1.00000000
```

```
> princomp(subjects)$loadings

Loadings:
        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
english -0.365 -0.363 -0.277  0.794 -0.166
math    -0.541  0.742 -0.393
socsci  -0.423 -0.325        -0.494 -0.686
natsci  -0.490         0.845  0.139  0.129
vocab   -0.393 -0.451 -0.231 -0.323  0.696

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
SS loadings       1.0    1.0    1.0    1.0    1.0
Proportion Var    0.2    0.2    0.2    0.2    0.2
Cumulative Var    0.2    0.4    0.6    0.8    1.0
```

PCA를 과목들에 적용해본 결과, 누적확률이 comp.3까지 약 90%이다.
따라서 각 과목들의 loading을 comp.3까지 살펴보자.
comp.1에서는 과목들의 전체적인 성적을 알 수 있다.
comp.2에서는 과학만 제외되어 있고, comp.3에서는 과학만 양의 값을 가진다.
다른 과목들에 비해 과학만 다른 양상을 보이므로 문과/이과 과목을 나누고자 한 목적을 이룰 수 없다.

또한 영어, 수학, 사회, 과학, 어휘 점수에 관한 상관관계를 보면 다음과 같다.

```
> cor(subjects)
          english      math    socsci    natsci     vocab
english 1.0000000 0.5412699 0.6339606 0.5783527 0.6692687
math    0.5412699 1.0000000 0.6086261 0.6583328 0.5556737
socsci  0.6339606 0.6086261 1.0000000 0.6749258 0.7719540
natsci  0.5783527 0.6583328 0.6749258 1.0000000 0.6004781
vocab   0.6692687 0.5556737 0.7719540 0.6004781 1.0000000
```

이 때 이과 과목인 수학은 과학과 상관관계가 높아야 하는데, 사회과목 역시 높은 편을 보이고 있다. 또한 우리의 데이터는 텍사스에서 조사한 데이터인데, 미국에서는 문과/이과를 나누고 있지 않다. 이러한 점들을 고려하여 과목을 특성에 따라 두 집단으로 나누는 데는 무리가 있다고 판단하였다. 그래서 5과목 모두의 평균으로 성적 데이터를 새로 정리하여 추가하여 분석하였다.

```
> head(twins)
  pairnum sex zygosity moed faed faminc english math socsci natsci vocab  even mean.sub
1       1   2        1    3    4      2      14   13     17     18    14 FALSE     15.2
2       1   2        1    3    4      2      11   14     15     10    12  TRUE     12.4
3       2   2        1    1    1      1      20   20     16     16    13 FALSE     17.0
4       2   2        1    1    1      1      17   19     13     13    14  TRUE     15.2
5       3   2        1    1    1      1      11    8     15     16    12 FALSE     12.4
6       3   2        1    1    1      1      16   13     13      8    15  TRUE     13.0
```

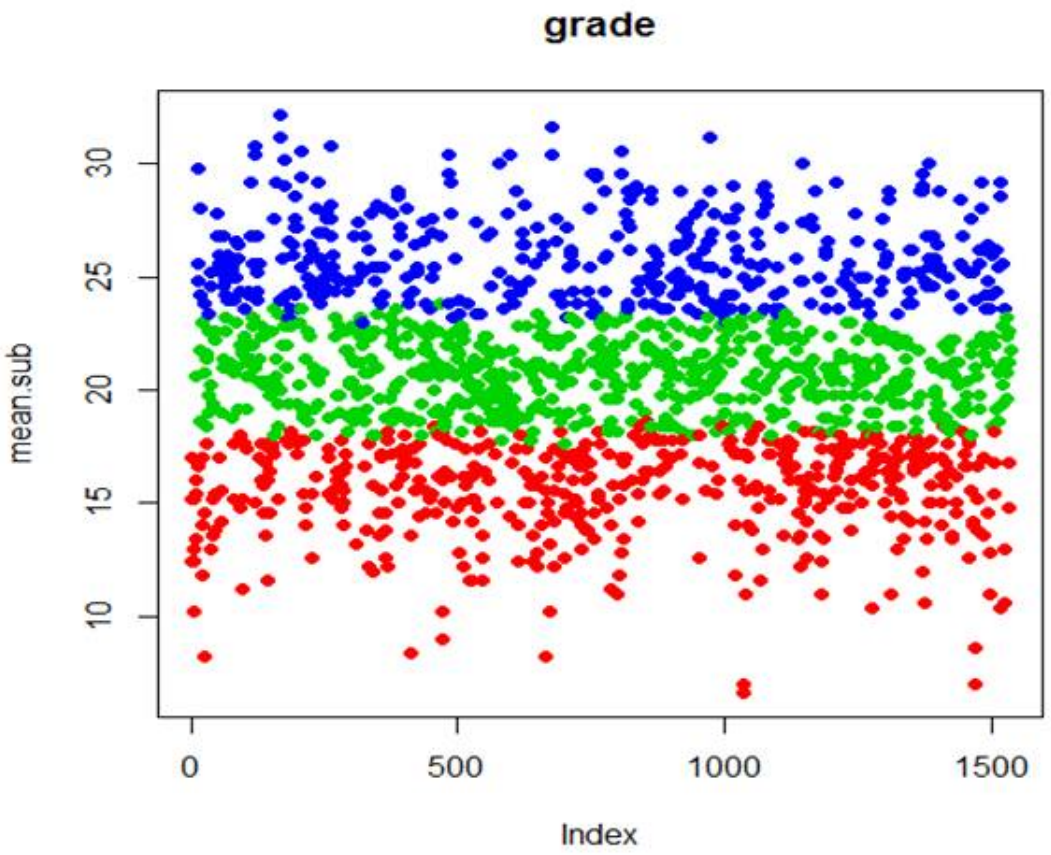
3) Clustering

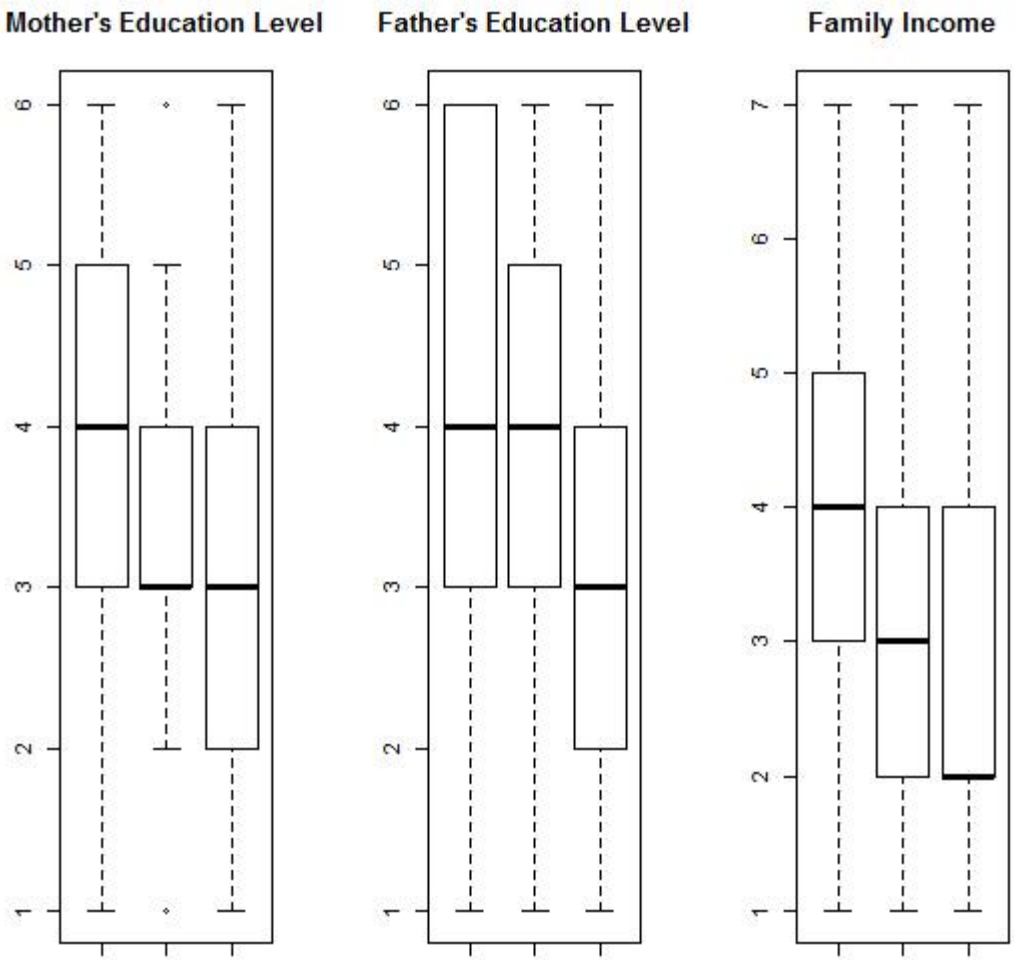
① 성적 간의 cluster (상, 중, 하)

과목들 간의 평균값에 대해서는 k.means 방법을 이용하여 집단을 나누었다. 상, 중, 하로 3집단으로 나누었고, 각각 476,627,433명의 학생들로 이루어져 있다. 그래프를 보다시피 각각의 평균은 약 15,20,25점 정도를 이루고 있다.

```
> kmeans.sub
K-means clustering with 3 clusters of sizes 476, 627, 433

Cluster means:
   english     math   socsci   natsci    vocab
1 15.68277 15.19118 15.91387 14.23529 16.54412
2 19.87241 21.27911 20.69378 20.55821 21.13876
3 24.07852 27.87529 25.94457 25.85450 25.92379
```





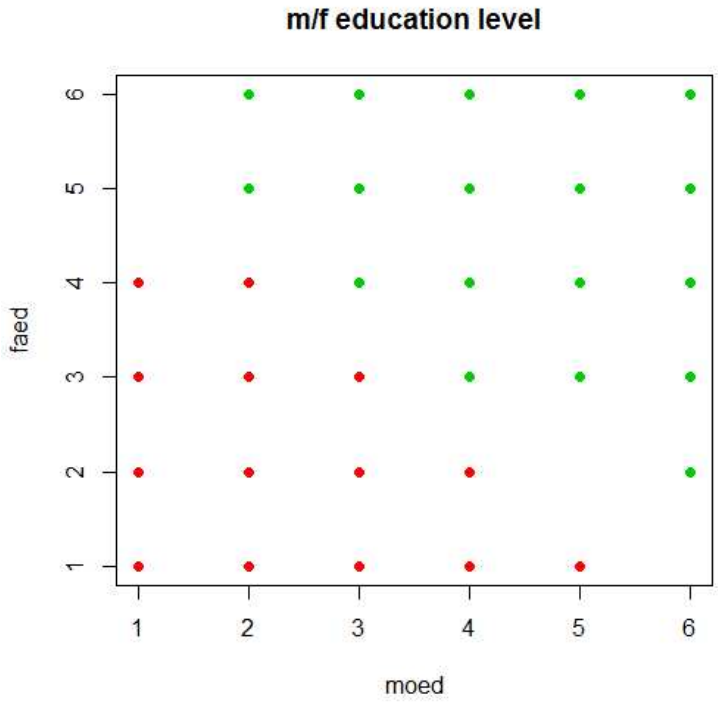
성적간의 어머니, 아버지의 학력수준과 가족 소득을 비교해보면, 성적이 높을수록 부모의 학력수준이 높고 소득수준도 높은 것으로 나타났다. 특히 아버지의 학력수준은 현저하게 영향력 있게 나타났다.

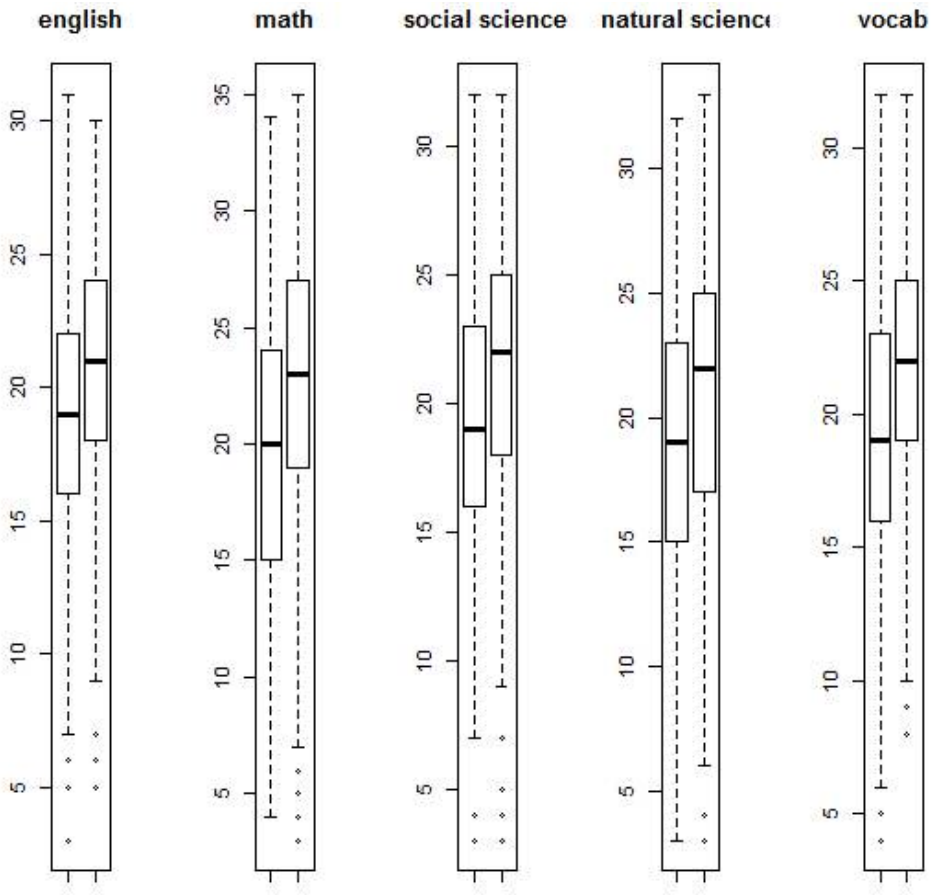
② 부모의 교육수준간의 cluster (고학력, 저학력)

어머니의 교육수준과 아버지의 교육수준간의 상관계수는 0.5582127 로 꽤 높은 편이므로, 부모의 교육수준을 평균값을 이용해 2개의 집단으로 나눴다. 고학력 집단은 439 가구로 평균적으로 대학교 졸업 이상의 학력이고, 저학력집단은 329 가구로 대학교를 졸업하지 않은 학력에 해당한다.

```
> kmeans.ed
K-means clustering with 2 clusters of sizes 658, 878

Cluster means:
      [,1]
1 4.756839
2 8.701595
```

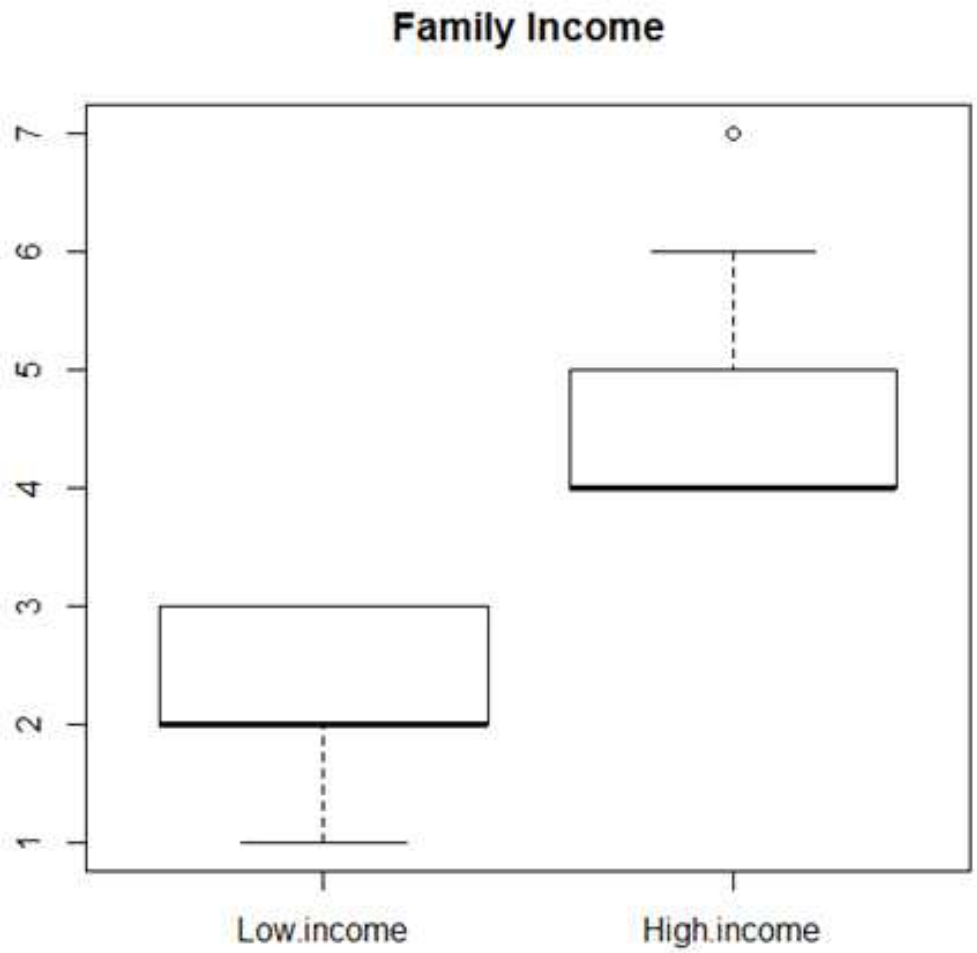




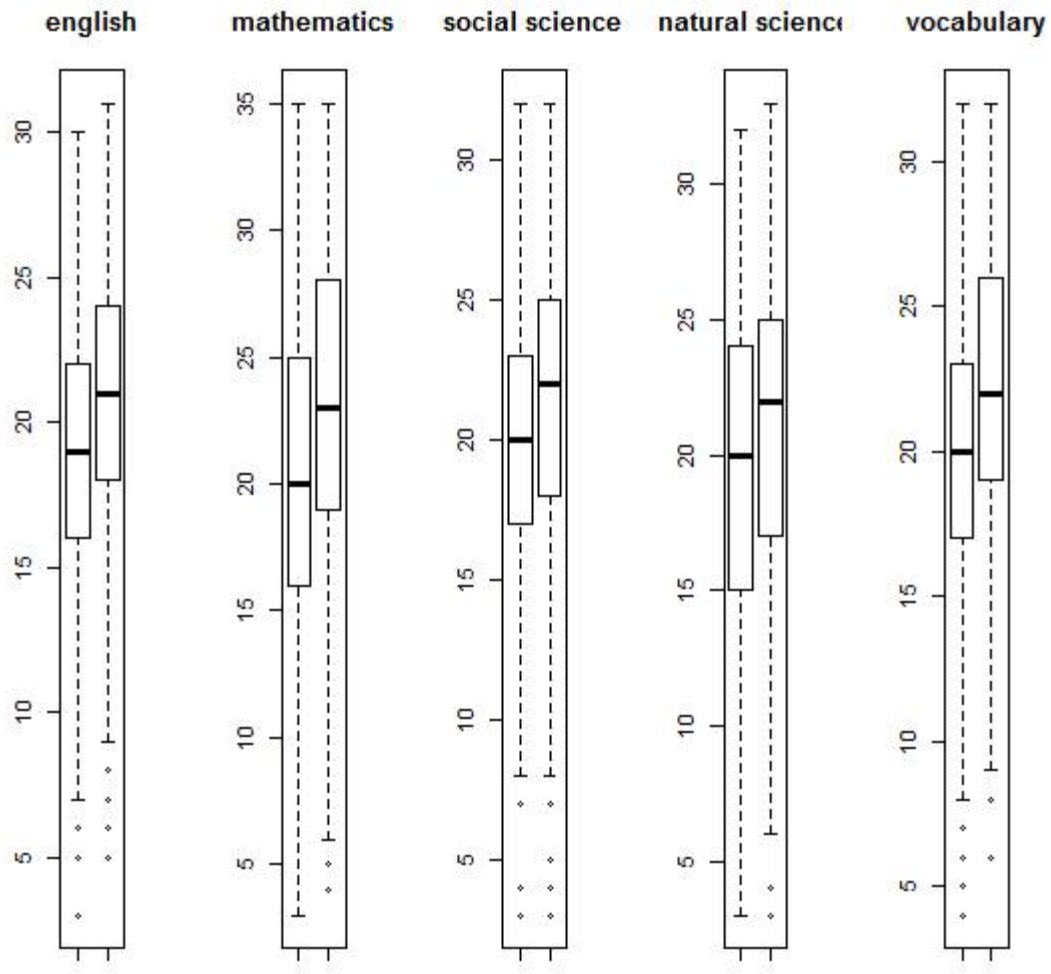
boxplot을 통해 살펴보면 모든 과목에서 고학력자 집안 자녀들이 높은 점수를 나타냈다. 특히 수학의 경우 성적의 차이가 컸다.

③ 가구의 소득 차에 따른 cluster (고소득, 저소득)

가구에 대한 소득수준 역시 k.means를 이용하여 집단을 2개로 나눴다. 그러나 원자료에서 연속변수인 소득 자료를 1부터 7로 이산 변수로 정리되어있기 때문에 간단히 자료 값이 3이하인 그룹을 inc1, 4이상인 그룹을 inc2로 나눌 수도 있다. 따라서 연소득이 10,000달러 이하인 가구를 저소득층, 10,000달러 이상인 가구는 고소득층으로 분류했다.



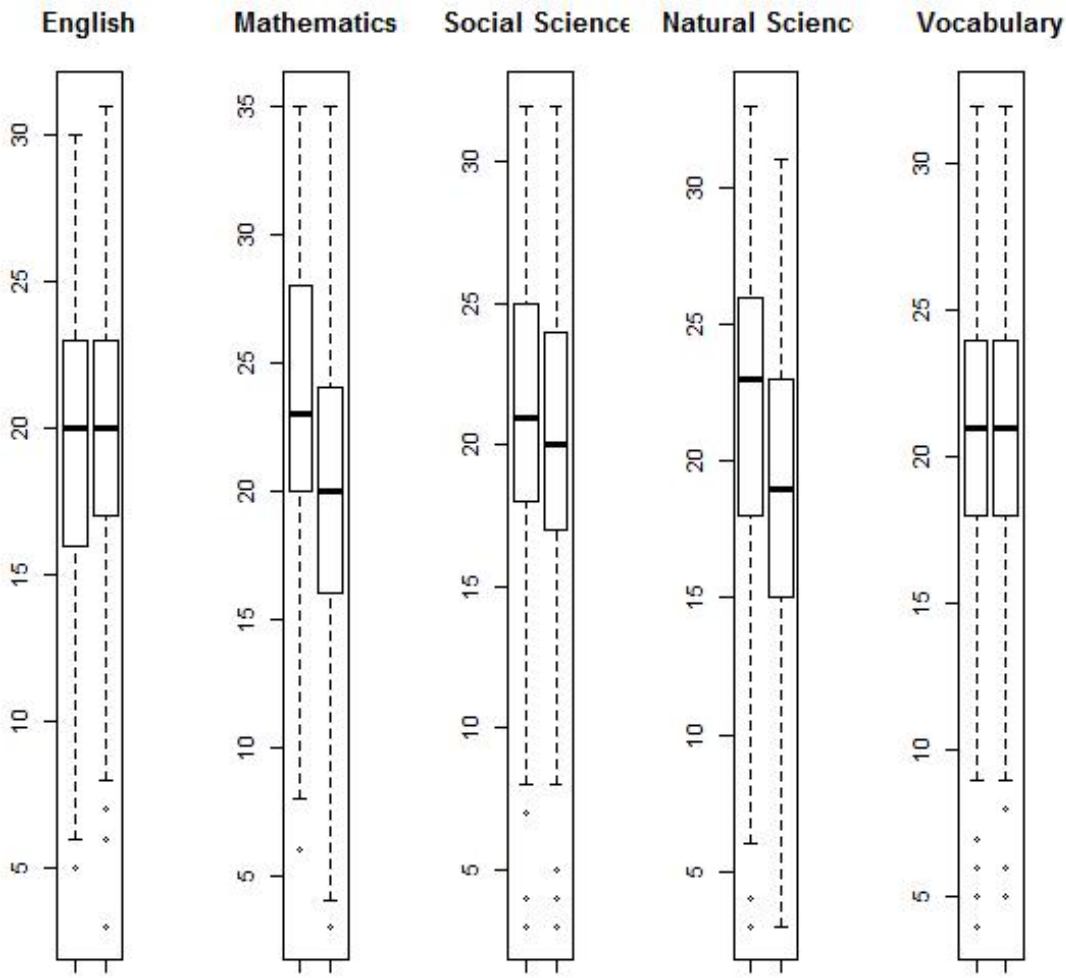
다음으로 소득별 과목의 점수 차이를 boxplot을 통해 분석했다.



소득수준에 따른 분포 역시 부모의 학력수준에서와 비슷하게 모든 과목에서 고소득층 자녀들이 높은 점수를 나타냈다. 또한 수학과 특히 어휘능력 부분에서 현저하게 높은 점수를 보였다.

4) 성별에 따른 점수 차이 비교

성별에 따른 과목간의 차이를 boxplot을 통해 비교해보았다. 영어와 어휘능력에 대해서는 성별에 대한 차이가 나타나지 않았다. 그러나 남학생들이 여학생에 비해 수학, 과학, 사회 3과목에 대해서 우수함을 나타냈다. 특히 과학과 수학에서는 남학생들이 눈에 띄게 높은 점수를 보이고 있다.



성별 간 과목간의 상관계수를 각각 살펴보면 아래와 같다. 이 때 특이한 점은 남성의 경우 가족의 소득과의 연관성이 높게 나타났고, 여성의 경우 아버지의 학력과 더 큰 관계를 보이고 있다.

```
> cor(cbind(male[,4:6],english))
              moed      faed     faminc    english
moed     1.0000000 0.5508148 0.4437020 0.1915980
faed     0.5508148 1.0000000 0.6370850 0.2290858
faminc   0.4437020 0.6370850 1.0000000 0.2177379
english  0.1915980 0.2290858 0.2177379 1.0000000
> cor(cbind(male[,4:6],math))
              moed      faed     faminc       math
moed     1.0000000 0.5508148 0.443702 0.2103076
faed     0.5508148 1.0000000 0.637085 0.2797060
faminc   0.4437020 0.6370850 1.000000 0.2963390
math     0.2103076 0.2797060 0.296339 1.0000000
> cor(cbind(male[,4:6],socsci))
              moed      faed     faminc     socsci
moed     1.0000000 0.5508148 0.4437020 0.1603984
faed     0.5508148 1.0000000 0.6370850 0.2158116
faminc   0.4437020 0.6370850 1.0000000 0.2243527
socsci   0.1603984 0.2158116 0.2243527 1.0000000
> cor(cbind(male[,4:6],natsci))
              moed      faed     faminc     natsci
moed     1.0000000 0.5508148 0.4437020 0.1884909
faed     0.5508148 1.0000000 0.6370850 0.2037480
faminc   0.4437020 0.6370850 1.0000000 0.2262221
natsci   0.1884909 0.2037480 0.2262221 1.0000000
> cor(cbind(male[,4:6],vocab))
              moed      faed     faminc      vocab
moed     1.0000000 0.5508148 0.4437020 0.2443759
faed     0.5508148 1.0000000 0.6370850 0.2770602
faminc   0.4437020 0.6370850 1.0000000 0.3107039
vocab    0.2443759 0.2770602 0.3107039 1.0000000

> cor(cbind(female[,4:6],english))
              moed      faed     faminc    english
moed     1.0000000 0.5640613 0.3764427 0.1403705
faed     0.5640613 1.0000000 0.4511499 0.2318194
faminc   0.3764427 0.4511499 1.0000000 0.1650638
english  0.1403705 0.2318194 0.1650638 1.0000000
> cor(cbind(female[,4:6],math))
              moed      faed     faminc       math
moed     1.0000000 0.5640613 0.3764427 0.1863463
faed     0.5640613 1.0000000 0.4511499 0.2817694
faminc   0.3764427 0.4511499 1.0000000 0.2364141
math     0.1863463 0.2817694 0.2364141 1.0000000
> cor(cbind(female[,4:6],socsci))
              moed      faed     faminc     socsci
moed     1.0000000 0.5640613 0.3764427 0.1657139
faed     0.5640613 1.0000000 0.4511499 0.2780162
faminc   0.3764427 0.4511499 1.0000000 0.1913037
socsci   0.1657139 0.2780162 0.1913037 1.0000000
> cor(cbind(female[,4:6],natsci))
              moed      faed     faminc     natsci
moed     1.0000000 0.5640613 0.3764427 0.1662560
faed     0.5640613 1.0000000 0.4511499 0.1816899
faminc   0.3764427 0.4511499 1.0000000 0.1528937
natsci   0.1662560 0.1816899 0.1528937 1.0000000
> cor(cbind(female[,4:6],vocab))
              moed      faed     faminc      vocab
moed     1.0000000 0.5640613 0.3764427 0.2064462
faed     0.5640613 1.0000000 0.4511499 0.3182588
faminc   0.3764427 0.4511499 1.0000000 0.2245227
vocab    0.2064462 0.3182588 0.2245227 1.0000000
```

5) Hotelling's T^2 – test

```
> twins1<-subset(twins,even==FALSE)
> twins2<-subset(twins,even==TRUE)
> twins.diff<-twins1-twins2
> HotellingsT2(twins.diff[,7:11])

        Hotelling's one sample T2-test

data:  twins.diff[, 7:11]
T.2 = 2.5559, df1 = 5, df2 = 763, p-value = 0.02637
alternative hypothesis: true location is not equal to c(0,0,0,0,0)
```

even 변수를 이용해 같은 쌍의 쌍둥이를 구별하였다.
쌍둥이 간의 차이를 구한 뒤 Hotelling's T^2 - test를 한 결과 p-value가 0.05보다 작은 것을 알 수 있다.
따라서 귀무가설을 기각한다.
이를 통해 같은 쌍의 쌍둥이 간에 성적의 차이가 있다고 결론 내릴 수 있다.

```
> t.test(english)

        One Sample t-test

data:  english
t = -2.1499, df = 767, p-value = 0.03187
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.54552508 -0.02478742
sample estimates:
 mean of x
-0.2851562
```

```
> t.test(math)

        One Sample t-test

data:  math
t = -2.4464, df = 767, p-value = 0.01465
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.84254826 -0.09234758
sample estimates:
 mean of x
-0.4674479
```

```
> t.test(socsci)

        One Sample t-test

data:  socsci
t = -2.543, df = 767, p-value = 0.01119
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.6414095 -0.0825488
sample estimates:
 mean of x
-0.3619792

> t.test(natsci)

        One Sample t-test

data:  natsci
t = -2.4171, df = 767, p-value = 0.01588
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.76686165 -0.07949252
sample estimates:
 mean of x
-0.4231771

> t.test(vocab)

        One Sample t-test

data:  vocab
t = -2.7065, df = 767, p-value = 0.006951
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.53017421 -0.08440912
sample estimates:
 mean of x
-0.3072917
```

모든 과목 성적에 대해 각각 t-test한 결과 p-value가 모두 0.05보다 작은 것을 알 수 있다.
따라서 귀무가설을 기각한다.
이를 통해 같은 쌍의 쌍둥이 간에 모든 과목들의 성적의 차이가 있다고 결론 내릴 수 있다.

5. 결론

1) 과목들을 PCA한 결과 과학만 다른 양상을 보이므로 문과/이과 과목으로 나누어 분석할 수 없다.

2) 성적 상/중/하 그룹을 나누어 분석한 결과 성적이 높을수록 부모의 학력수준과 소득수준이 높은 것으로 나타났다.

3) 부모의 교육수준을 상/하 그룹으로 나누어 분석한 결과 교육수준이 높을수록 모든 과목에서 성적이 높은 것으로 나타났다. 특히 수학의 경우 성적의 차이가 컸다.

4) 가정의 소득수준을 상/하 그룹으로 나누어 분석한 결과 소득수준이 높을수록 모든 과목에서 성적이 높은 것으로 나타났다. 특히 수학과 어휘의 경우 성적의 차이가 컸다.

5) 남/여 성별 그룹으로 나누어 분석한 결과 영어와 어휘의 경우 성적의 차이가 없었고 수학, 과학, 사회의 경우 남성 그룹이 여성 그룹보다 성적이 높았다.
남성 그룹의 경우 소득 수준과, 여성 그룹의 경우 아버지의 학력과 더 큰 관계를 보였다.

6) 같은 쌍의 쌍둥이 간에 전체적인 성적뿐만 아니라 각각의 과목에 대한 성적에서도 차이가 나타났다.

자료출처 :
http://psych.colorado.edu/~carey/Courses/PSYC7291/ClassDataSets.htm