R_HW_6_Multilevel Models

Eom SangJun

2020 11 2

Multilevel model 은 계층적 구조를 가진 데이터에 대한 모델을 지칭한다.

학생들을 대상으로 math 성적이나 gender, social class of father 등을 조사한 데이터를 이용하여 이를 알아보자.

```
data(jsp, package = 'faraway')
jspr <- jsp[jsp$year==2,]
```

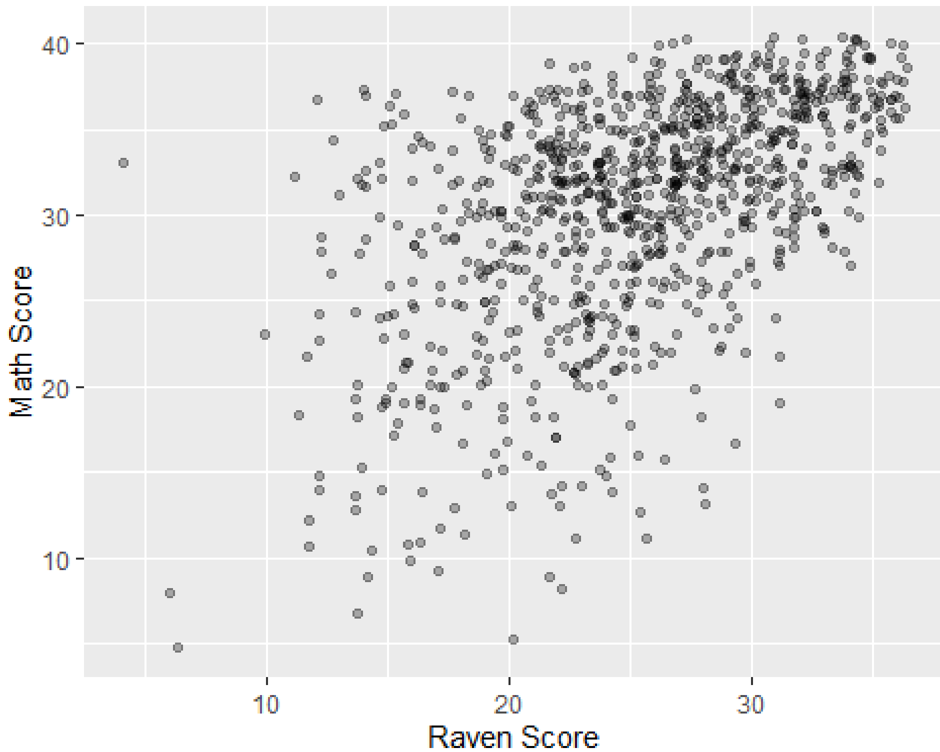
→ 우리는 final year 의 math test score 를 종속변수로 사용할 것이기 때문에 final year 만을 남긴다.

Raven’s test score 는 입학 당시의 학생의 능력을 평가하기 위한 시험의 성적으로써 이를 종속변수인 math test score 와 비교해보자.

```
library(ggplot2)
```

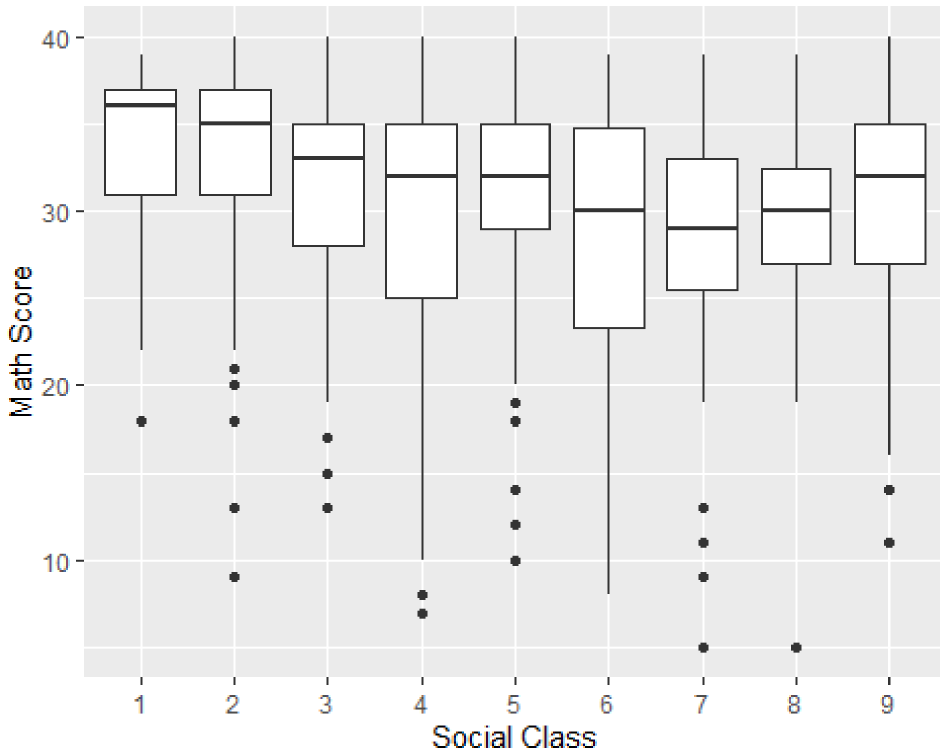
```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
ggplot(jspr, aes(x=raven, y=math)) +
  xlab('Raven Score') +
  ylab('Math Score') +
  geom_point(position=position_jitter(), alpha=0.3)
```



➔ 대략적으로 양의 상관관계를 보인다는 것을 알 수 있다.

```
ggplot(jspr, aes(x=social, y=math)) +
  xlab('Social Class') +
  ylab('Math Score') +
  geom_boxplot()
```



➔ Social class 와 math test score 간의 관계를 살펴보면 class 가 높을수록(1 에 가까울수록 높다)math test score 도 어느 정도 높다는 것을 알 수 있다.

현재 데이터를 분석하는 방법 중 하나는 multiple regression 이다.

```
glin <- lm(math ~ raven*gender*social, jspr)
anova(glin)
```

```
## Analysis of Variance Table
##
## Response: math
##                      Df  Sum Sq Mean Sq  F value    Pr(>F)
## raven                 1 11480.5 11480.5 368.0625 < 2.2e-16 ***
## gender                1    44.1    44.1   1.4142  0.234668
## social                8   779.4    97.4   3.1233  0.001725 **
## raven:gender          1     0.0     0.0   0.0004  0.984718
## raven:social          8   582.6    72.8   2.3347  0.017460 *
## gender:social         8   450.1    56.3   1.8038  0.072742 .
## raven:gender:social   8   234.6    29.3   0.9400  0.482355
## Residuals           917 28602.8    31.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

→ gender 가 포함된 변수들은 모두 유의미하지 않다는 결과를 얻었다. 따라서 제외하고 다시 모델링을 해보자.

```
glin <- lm(math ~ raven*social, jspr)
anova(glin)
```

```
## Analysis of Variance Table
##
## Response: math
##               Df  Sum Sq Mean Sq  F value    Pr(>F)
## raven          1 11480.5 11480.5 365.7151 < 2.2e-16 ***
## social         8   777.6    97.2   3.0964  0.001869 **
## raven:social   8   564.5    70.6   2.2477  0.022241 *
## Residuals    935 29351.5    31.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

→ 우리가 가진 데이터는 꽤 큰 데이터이기 때문에 심지어 조그만 effect 도 유의미하다고 나올 수 있다. 따라서 raven:social 이 비록 유의미하다고 나오긴 했지만, 해석의 편의성을 위해 제거해주도록 하자.

```
glin <- lm(math ~ raven + social, jspr)
summary(glin)
```

```
##
## Call:
## lm(formula = math ~ raven + social, data = jspr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.8430  -3.2426   0.7726   3.7765  14.0825
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.02481    1.37451  12.386   <2e-16 ***
## raven        0.58040    0.03256  17.826   <2e-16 ***
## social2      0.04950    1.12938   0.044   0.9651
## social3     -0.42893    1.19568  -0.359   0.7199
## social4     -1.77452    1.05993  -1.674   0.0944 .
## social5     -0.78228    1.18924  -0.658   0.5108
## social6     -2.49373    1.26094  -1.978   0.0483 *
## social7     -3.04851    1.29065  -2.362   0.0184 *
## social8     -3.11746    1.77494  -1.756   0.0793 .
## social9     -0.63278    1.12731  -0.561   0.5747
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.632 on 943 degrees of freedom
```

```
## Multiple R-squared:  0.2907, Adjusted R-squared:  0.2839
## F-statistic: 42.93 on 9 and 943 DF,  p-value: < 2.2e-16
```

→ 우리는 final math score 가 Raven score 와 매우 강한 상관관계를 가지고 있다는 것을 알 수 있으며 social class 가 낮을수록 math score 도 낮은 경향을 보인다는 것을 알 수 있다.

그런데 이러한 multiple regression 에는 문제점이 있다. 바로 학생들이 모두 독립적이라고 가정한 것이다. 하지만 우리는 쉽게 같은 학교에서 온 학생들은 어느 정도 dependency 가 있을 것이라고 쉽게 생각할 수 있다. 만약 dependency 가 있음에도 불구하고 없다고 가정하는 경우 결과의 significance 를 overstate 할 가능성이 존재한다. 더 나아가 위의 분석방식은 학교 간의 그리고 학교 내의 variation 을 설명할 수 없으며 학교 간의 데이터를 단순 통합해서 처리해버리는 방식은 정보의 손실을 가져올 수 있다. 따라서 우리는 학생 개별 수준의 정보를 이용하면서도, 데이터 내의 grouping 을 반영해줄 수 있는 다른 분석 방식을 사용하고자 한다.

각 학교에서 온 학생들의 수의 표는 다음과 같다.

```
table(jspr$school)
```

```
## 
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25 
 26 
##  26  11  14  24  26  18  11  27  21   0  11  23  22  13   7  16   6  18  14  13  28  14  18  21  14 
 20 
##  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  44  45  46  47  48  49  50 
##  22  15  13  27  35  23  44  27  16  28  17  12  14  10  10  41   5  11  15  33  63  22  14
```

```
library(lme4)
```

```
## Warning: package 'lme4' was built under R version 3.6.3
```

```
## Loading required package: Matrix
```

우선 우리는 단순 Linear Model 대신 Linear Mixed Model 을 fitting 할 것이며, fixed effect 로는 raven, social, gender 간의 모든 interaction 을, random effect 로는 school 과 class nested within the school 을 상정할 것이다.

```
#lmer --> Linear Mixed Model fitting
mmod <- lmer(math ~ raven*social*gender+(1|school)+(1|school:class), data=jsp
r)
```

→ 이번에도 마찬가지로, summary 값을 확인하면 gender 의 significance 값이 낮다는 것을 확인할 수 있는데, Kenward-Roger adjusted F-test 를 이용해서 gender variable 을 제외한 모델이 유의미한 지를 살펴볼 수 있다.

```
library(pbkrtest)
```

```
## Warning: package 'pbkrtest' was built under R version 3.6.3
```

```r
mmodr <- lmer(math ~ raven*social+(1|school)+(1|school:class), data=jspr)
KRmodcomp(mmod, mmodr)
```

```
## F-test with Kenward-Roger approximation; time: 1.22 sec
## large : math ~ raven * social * gender + (1 | school) + (1 | school:class)
## small : math ~ raven * social + (1 | school) + (1 | school:class)
##          stat      ndf      ddf F.scaling p.value
## Ftest   1.0137  18.0000 892.9395   0.99997   0.441
```

→ p-value 가 높은데 여기서는 귀무가설이 large 가 아니라 small 이다. 따라서 small model 을 택한다.

우리는 model selection 에 있어서 criterion-based approach 를 채택할 수 있는데, 그 중 하나로 우리가 고려하고 싶은 모든 모델을 특정하는 방법을 생각해볼 수 있다.

우리가 고려해보고자 하는 모델들은 다음과 같다.

```r
all3 <- lmer(math ~ raven*social*gender + (1|school) + (1|school:class),
             data=jspr, REML = FALSE)
all2 <- update(all3, . ~ . - raven:social:gender)
notrs <- update(all2, . ~ . - raven:social)
notrg <- update(all2, . ~ . - raven:gender)
notsg <- update(all2, . ~ . - social:gender)
onlyrs <- update(all2, . ~ . - social:gender - raven:gender)
all1 <- update(all2, . ~ . - social:gender - raven:gender - social:raven)
nogen <- update(all1, . ~ . -gender)
```

그리고 anova function 을 이용해서 AIC 와 BIC 를 구해보면 다음과 같다.

```r
anova(all3, all2, notrs, notrg, notsg, onlyrs, all1, nogen)[,1:4]
```

```
##        npar    AIC    BIC  logLik
## nogen    13 5954.3 6017.5 -2964.2
## all1     14 5955.6 6023.6 -2963.8
## onlyrs   22 5950.1 6057.0 -2953.1
## notrs    23 5961.6 6073.4 -2957.8
## notsg    23 5952.0 6063.8 -2953.0
## notrg    30 5956.1 6101.9 -2948.1
## all2     31 5957.8 6108.4 -2947.9
## all3     39 5966.7 6156.2 -2944.3
```

→ 원래 anova function 은 model 들을 비교하기 위해 chi-squared test 를 한다. 그러나 model 들이 nested 되어있지 않기 때문에 여기서는 적절하지 않으며, different fixed effect 를 가진

model 들을 비교하는 데에 REML Method 는 부정확하다. 따라서 [,1:4]를 이용하여 필요 없는 부분은 제거해주었다.

결과적으로 gender 를 제외한 nogen model 의 AIC 가 제일 낮다는 것을 확인할 수 있다.

```
library(faraway)
```

```
## Warning: package 'faraway' was built under R version 3.6.3
```

Gender 가 중요하지 않다는 것을 알았으니 gender 를 제외하고 다시 modeling 을 해보자.

```
jspr$craven <- jspr$raven - mean(jspr$raven)
mmod <- lmer(math ~ craven*social+(1|school)+(1|school:class), jspr)
sumary(mmod)
```

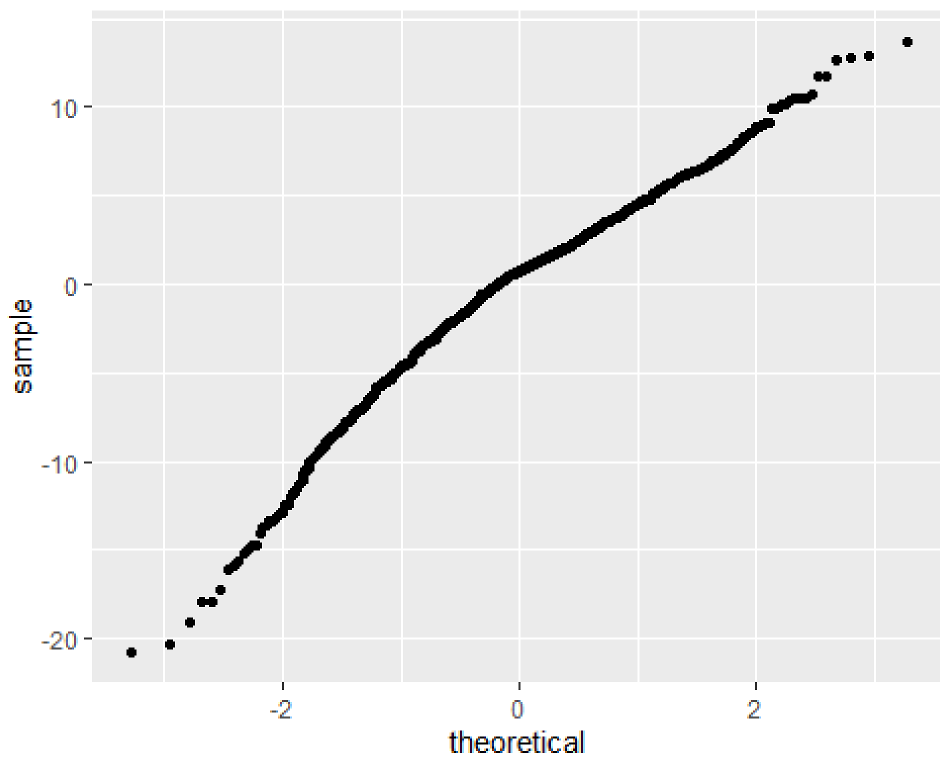
```
## Fixed Effects:
##                coef.est coef.se
## (Intercept)    31.91     1.20
## craven          0.61     0.19
## social2         0.02     1.27
## social3        -0.63     1.31
## social4        -1.97     1.20
## social5        -1.36     1.30
## social6        -2.27     1.37
## social7        -2.55     1.41
## social8        -3.39     1.80
## social9        -0.83     1.25
## craven:social2 -0.13     0.21
## craven:social3 -0.22     0.22
## craven:social4  0.04     0.19
## craven:social5 -0.15     0.21
## craven:social6 -0.04     0.23
## craven:social7  0.40     0.23
## craven:social8  0.26     0.26
## craven:social9 -0.08     0.21
##
## Random Effects:
##  Groups       Name        Std.Dev.
##  school:class (Intercept) 1.08
##  school       (Intercept) 1.77
##  Residual                 5.21
## ---
## number of obs: 953, groups: school:class, 90; school, 48
## AIC = 5963.2, DIC = 5893.6
## deviance = 5907.4
```

→ Modeling 을 하기에 앞서 우선 Raven score 를 center 화 시켜주었다. 이는 social 에 따른 차이를 비교할 때 raven test score 가 0 일 때가 아니라 mean 일 때 social effect 를 확인하는 것이 더 적절하기 때문이다.

분석 결과 Raven score 는 math score 와 강한 상관관계를 가지고 있으며, class 가 낮을수록 math score 가 낮은 경향을 보인다. 그러나 class 의 경우 9 가 8 이나 7 에 비해 더 낮지 않은 것을 보았을 때 완전히 ordinal 은 아니다.

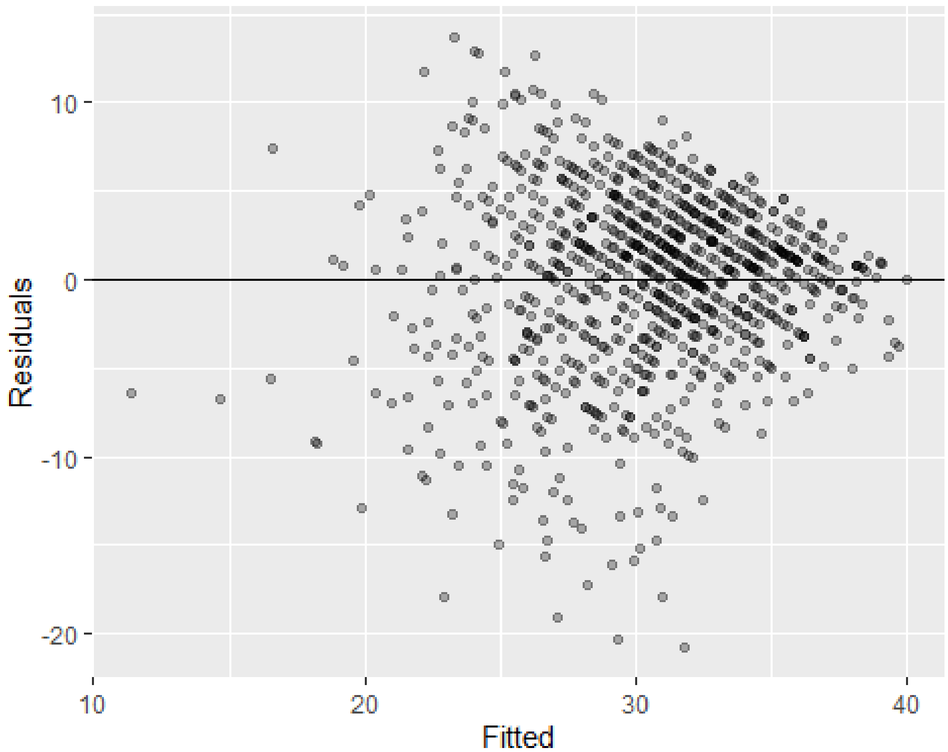
또한 데이터 간의 대부분의 variation 은 individual level 에서 오며 variation at the school 과 class level 은 그것보다 작다.

```
diagd <- fortify(mmod)
ggplot(diagd, aes(sample=.resid))+stat_qq()
```



➔ QQ plot 에는 문제가 없는 것으로 보인다.

```
ggplot(diagd, aes(x=.fitted, y=.resid)) + geom_point(alpha=0.3) +
  geom_hline(yintercept = 0) + xlab('Fitted') + ylab('Residuals')
```

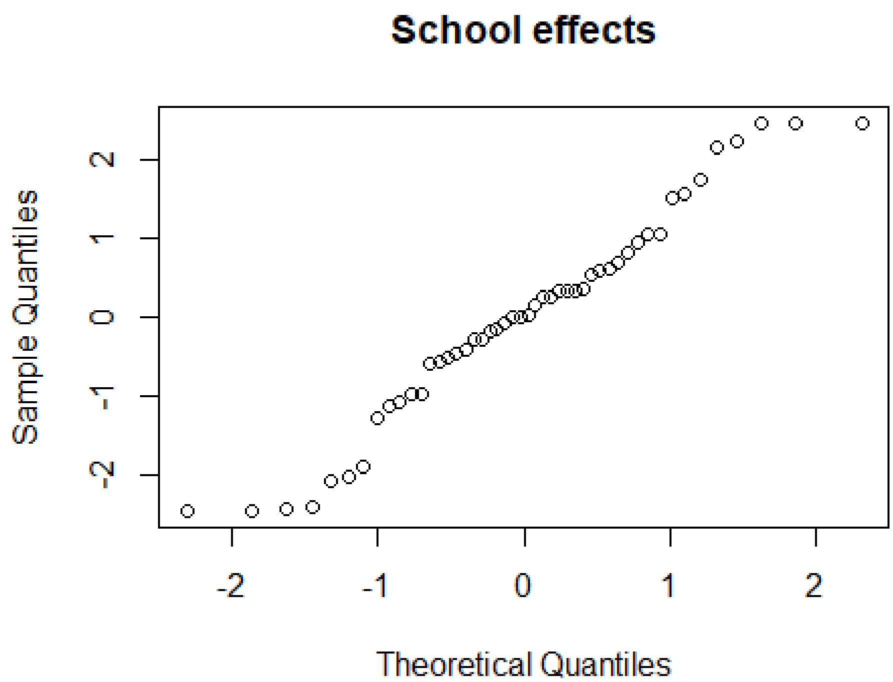


➔ Fitted value 가 증가할수록 variance 가 줄어드는 경향성을 보인다.

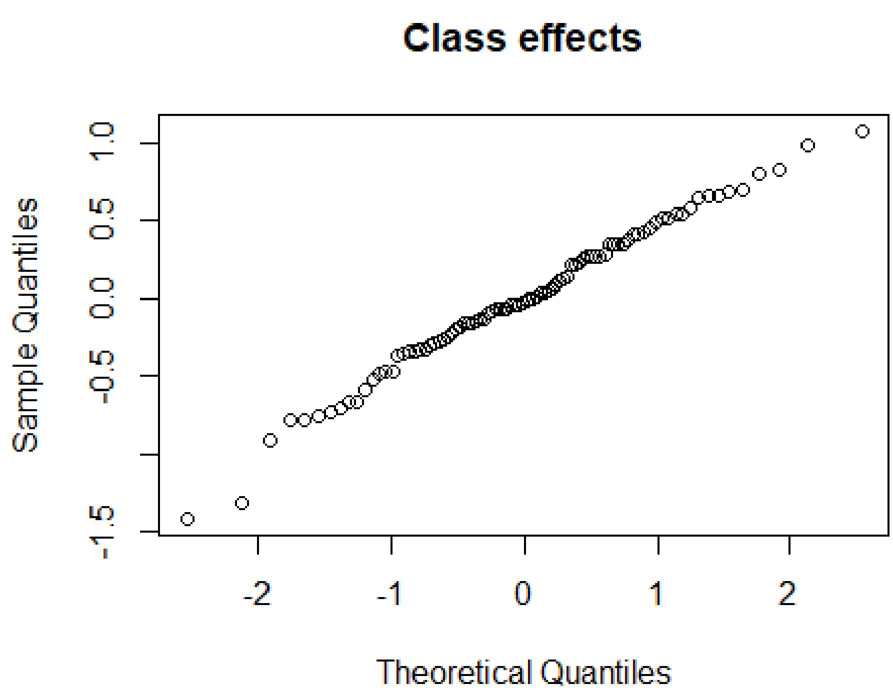
➔ 이는 최대 총점이 40 점으로 제한되어 있기 때문에 나타나는 문제점이라고 볼 수 있다.

➔ 우리는 이를 해결하기 위해 종속변수의 transformation 을 고려해야 한다.

```
qqnorm(ranef(mmod)$school[[1]], main='School effects')
```




```
qqnorm(ranef(mmod)$'school:class'[[1]], main = 'Class effects')
```



➔ Random effect 들도 normally distributed 되어 있다는 것을 확인할 수 있다.

흥미롭게도 우리는 school effect 에 대해 좀 더 자세히 살펴볼 수 있다.

비록 학생들의 최종 math score 가 높은 것으로 보이더라도 만약 원래 학생들의 성적이 좋은 학교였다면 school effect 는 그다지 크지 않을 수 있다. 또한 최종 성적이 높더라도 입학 성적에 비해 그것이 감소한 것이라면, 오히려 그 학교는 부정적인 effect 를 가지고 있다는 것을 확인할 수 있을 것이다. 이를 살펴보자.

우선 quality of intake 와 학생들의 class 를 고려한 school 들의 math score ranking 을 다음과 같이 구하자.

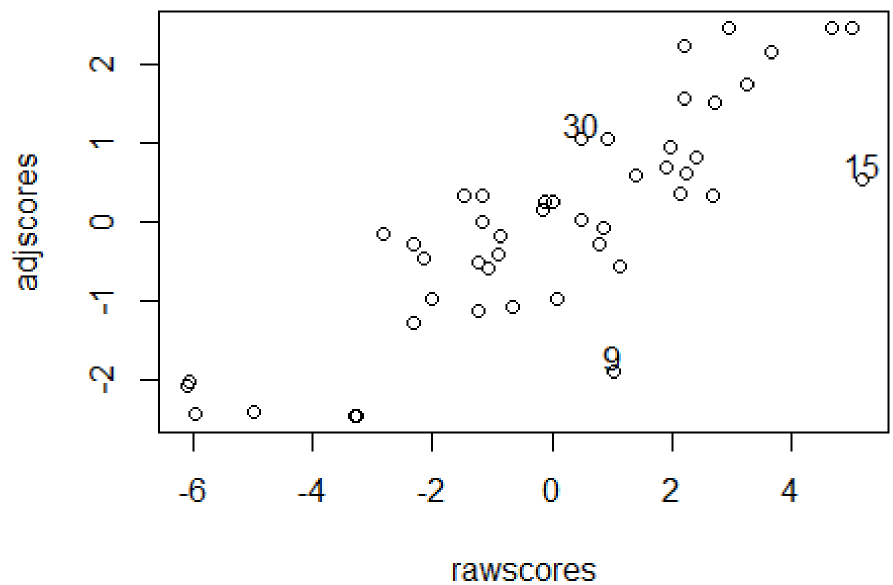
```
adjscores <- ranef(mmod)$school[[1]]
```

다음으로는 adjusted 되지 않는 raw score 또한 구해보자.

```
rawscores <- coef(lm(math ~ school-1, jspr))
rawscores <- rawscores - mean(rawscores)
```

그 다음 두 점수를 비교하면 다음과 같다.

```
plot(rawscores, adjscores)
sint <- c(9, 14, 29) #school 10 은 list 에 있긴 하지만, 학생 수가 0 명이므로 이를
 반영해주어야 한다.
text(rawscores[sint], adjscores[sint]+0.2, c('9', '15', '30'))
```



➔ 세 학교가 눈에 띄는 결과가 나왔다.

➔ 학교 30 의 경우 raw score 는 별로 높지 않지만 intake 와 social class 를 고려한 결과 더 높은 score 를 보인다.

➔ 반면 15 와 9 의 경우 raw score 는 높지만 intake 와 social class 를 고려해보면 별로 좋지 않은 결과를 얻었다는 것을 알 수 있다.

우리는 학교 간의 또는 class 간의 정말로 variation 이 얼마나 있는 지에 대해 관심이 있을 수 있다. 이를 알아보기 위해 다음과 같은 방식을 사용한다.

```
library(RLRsim)

## Warning: package 'RLRsim' was built under R version 3.6.3

mmodc <- lmer(math ~ craven*social+(1|school:class), jspr)
mmods <- lmer(math ~ craven*social+(1|school), jspr)

exactRLRT(mmodc, mmod, mmods)

##
##  simulated finite sample distribution of RLRT.
##
```

```
##  (p-value based on 10000 simulated values)
##
## data:
## RLRT = 2.3903, p-value = 0.0538
```

→ class effect 는 통계적으로 유의미한 경계에 있다는 것을 알 수 있다. 만약 fixed effect term testing 을 위해 고려한다고 하더라도 그 영향은 별로 크지 않다는 것을 알 수 있다.

```
exactRLRT(mmods, mmod, mmodc)
```

```
##
##  simulated finite sample distribution of RLRT.
##
##  (p-value based on 10000 simulated values)
##
## data:
## RLRT = 7.1403, p-value = 0.0034
```

→ 반면 school effect 는 굉장히 유의미하며 class 의 것보다 더 크다는 것을 알 수 있다. 즉, 특정 선생보다 특정 학교가 중요하다는 것을 알 수 있다.

우리가 지금까지 살펴본 fixed effect 들은 모두 individual 수준에서의 것들이었다. 그런데 school 이나 class level 의 fixed effect 도 우리는 고려해볼 수 있는데 이를 compositional effect 라고 부른다. 예를 들어 학교 친구들의 성적은 어떤 학생에게 큰 영향을 미칠 것이라고 가정할 수 있다. 즉, 어떤 학교의 평균적인 입학성적은 개개인의 최종 math score 에 영향을 미칠 수 있다고 가정할 수 있다. 따라서 이를 반영해줄 수 있는 variable 을 만들어보면 다음과 같다.

```
schraven <- lm(raven ~ school, jspr)$fit
```

```
mmodc <- lmer(math ~ craven*social+schraven*social+(1|school)+(1|school:clas
s), jspr)
KRmodcomp(mmod, mmodc)
```

```
## F-test with Kenward-Roger approximation; time: 0.79 sec
## large : math ~ craven * social + schraven * social + (1 | school) + (1 |
##     school:class)
## small : math ~ craven * social + (1 | school) + (1 | school:class)
##          stat     ndf      ddf F.scaling p.value
## Ftest   0.6789   9.0000 640.1393   0.99707  0.7285
```

→ 아쉽게도 F-test 결과 새로운 variable 은 유의미하지 않다는 결과가 나왔다. 우리는 단순히 평균만을 고려하였는데 quantile 이나 다른 spread measure 방법등을 고려하여 이를 variable 에 반영해주는 것도 가능하다.