Homework 7

Statistical Inference II

Noah Johnson

April 19, 2018

In this homework, the data based on a sample of n = 176 chidren within J = 10 schools in the American subsample of the PISA (Programme for International Student Assessment) and is available in PISASchools10.sav. It has intentionally been provided in an .sav format so you will have to find the function to read a .sav file.

First, we will look at the relationship between student's home education resources (HEDRES) on math achievement scores (MATHSCOR) across these 10 schools. Then, we will examine whether the schools' press for academic excellence (ACADPRES) moderates this relationship. Read each question carefully, as there are multiple parts to most questions.

HEDRES scaled from 1 to 8; lower values indicate poor home resources for education

MATHSCOR average of 50; SD of 9 points

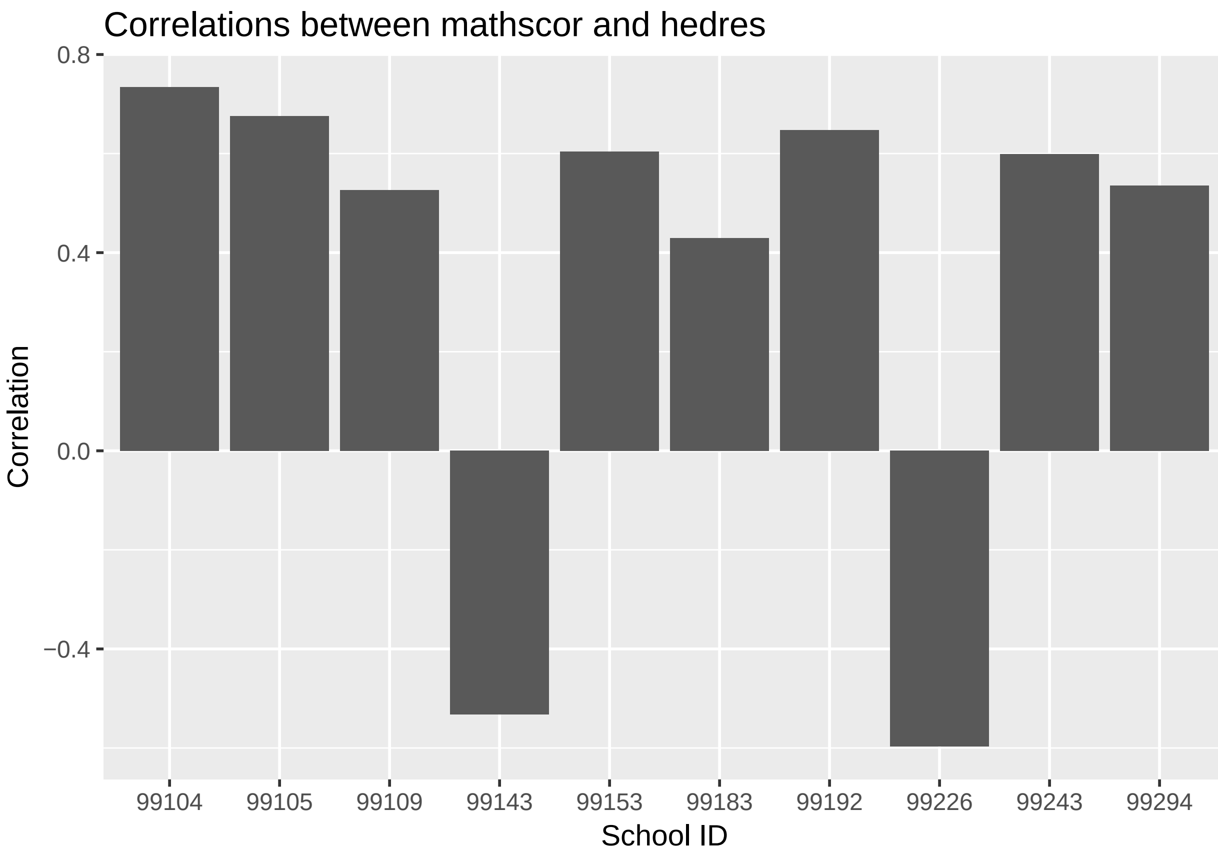
ACADPRES scaled from 1 to 8; lower values indicate low press for academic excellence

1. Use the *sav* data file to run separate multiple regression models for each of the 10 schools (X = *hedres*; Y = *mathscor*). Review the separate regression results for each of the schools. What do you notice about the results (look at the correlations between *mathscor* and *hedres*, and the regression coefficients for each of the 10 schools)? Is it reasonable to assume that the effect of home resources on math achievement is the same in all 10 schools? Why or why not? What does your answer imply regarding how to include the *hedres* variable within our HLM?

```
# Read in data
data <- read_sav('PISASchools10.sav')

# Compute 10 different linear models
lms <- data %>%
  group_by(schoolid) %>%
  do(fit = lm(data = ., formula = mathscor ~ hedres), correlation = cor(.$mathscor, .$hedres)) %>%
  mutate(correlation = unlist(correlation))

# Look at correlations
lms %>% select(schoolid, correlation) %>% ggplot(aes(x=factor(schoolid), y = correlation)) +
  geom_bar(stat = 'identity') +
  labs(x = 'School ID', y = 'Correlation', title = 'Correlations between mathscor and hedres')
```



Schools 99143 and 99226 have a negative relationship between mathscor and hedres, while the other schools have a positive relationship.

```
# Look at regression coefficients
tidy(lms, fit)
```

```
## # A tibble: 20 x 7
## # Groups:   schoolid, correlation [10]
##    schoolid correlation        term  estimate  std.error  statistic
##       <dbl>       <dbl>       <chr>     <dbl>      <dbl>      <dbl>
##  1    99104   0.7338627 (Intercept) 27.905634  7.9028023  3.5311062
##  2    99104   0.7338627      hedres  4.074789  1.1372455  3.5830339
##  3    99105   0.6752665 (Intercept) 29.194281  4.4796420  6.5171015
##  4    99105   0.6752665      hedres  2.612235  0.7133171  3.6620955
##  5    99109   0.5265970 (Intercept) 27.920323  6.8027142  4.1042916
##  6    99109   0.5265970      hedres  2.639680  1.0653457  2.4777684
##  7    99143  -0.5323447 (Intercept) 85.652684 12.6979107  6.7454155
##  8    99143  -0.5323447      hedres -4.618533  1.8360827 -2.5154274
##  9    99153   0.6036368 (Intercept)  8.859310 11.5459855  0.7673065
## 10    99153   0.6036368      hedres  5.241953  1.7876048  2.9323894
## 11    99183   0.4290468 (Intercept) 30.300841 14.3337318  2.1139534
## 12    99183   0.4290468      hedres  3.627537  1.9719004  1.8396145
## 13    99192   0.6474948 (Intercept) 25.516067  5.2914473  4.8221338
## 14    99192   0.6474948      hedres  2.867856  0.7955723  3.6047715
## 15    99226  -0.5968380 (Intercept) 70.317902  5.9505211 11.8170998
## 16    99226  -0.5968380      hedres -3.076326  1.0030479 -3.0669785
## 17    99243   0.5992993 (Intercept) 23.350127  8.8541099  2.6372077
## 18    99243   0.5992993      hedres  4.539951  1.5160805  2.9945315
```

```
## 19    99294   0.5350646 (Intercept) 28.789028 10.3401905  2.7841874
## 20    99294   0.5350646      hedres  3.748527  1.4796332  2.5334165
## # ... with 1 more variables: p.value <dbl>
```

Here we see the same split, with two models having a negative slope. Clearly the effect of home resources on math achievement is NOT the same in all 10 schools. Thus Hedres should be included as a Level 1 effect within our HLM.

2. MODEL 1: Using lme4 fit an unconditional random-effects ANOVA (i.e., *empty model*) with *mathscor* as the outcome. Report and interpret all the parameters as well as the ICC.

```
model.empty <- lmer(mathscor ~ 1 + (1|schoolid), data = data)
model.empty.sum <- summary(model.empty)
model.empty.sum
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: mathscor ~ 1 + (1 | schoolid)
##    Data: data
##
## REML criterion at convergence: 1232.9
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.29327 -0.66535  0.01497  0.65333  2.23111
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  schoolid (Intercept) 27.13    5.209
##  Residual             58.19    7.628
## Number of obs: 176, groups:  schoolid, 10
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   49.883      1.746   28.57
```

$mathscor_{ij} = \beta_{0j} + r_{ij}$

$\beta_{0j} = \gamma_{00} + u_{0j}$

$\gamma_{00} = 49.8825915$ is the estimated grand mean math score.

$var(r_{ij}) = 58.19 = s^2$ is the within-group variance.

$var(u_{0j}) = 27.13$ is the between-group variance.

$ICC = \frac{27.13}{27.13+58.19} = 0.3179794$ is the intraclass correlation coefficient. Since it is not zero, some variance in mathscor is accounted for by schoolid.

3. MODEL 2: Run a *random coefficients model* with *hedres* (**group-mean centered**) as the predictor of math achievement. Report and interpret all the parameters. **Compared to Model 1**, how much was the within-schools variability (s^2) reduced with the addition of the group-centered home resources variable?

```
school_group_means <- data %>%
  group_by(schoolid) %>%
  summarize(gpm_hedres = mean(hedres))

data <- merge(data, school_group_means, by = "schoolid")
data$hedres.group.mean.cen <- data$hedres - data$gpm_hedres
```

```r
model2 <- lmer(mathscor ~ hedres.group.mean.cen + (hedres.group.mean.cen|schoolid),
               data = data)
model2.sum <- summary(model2)
model2.sum
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## mathscor ~ hedres.group.mean.cen + (hedres.group.mean.cen | schoolid)
##    Data: data
##
## REML criterion at convergence: 1186.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.6170 -0.6908  0.1315  0.6730  2.0353
##
## Random effects:
##  Groups   Name                  Variance Std.Dev. Corr
##  schoolid (Intercept)           28.111   5.302
##           hedres.group.mean.cen  9.101   3.017    -0.40
##  Residual                       40.391   6.355
## Number of obs: 176, groups:  schoolid, 10
##
## Fixed effects:
##                       Estimate Std. Error t value
## (Intercept)             49.882      1.744  28.595
## hedres.group.mean.cen    1.937      1.056   1.834
##
## Correlation of Fixed Effects:
##             (Intr)
## hdrs.grp.m. -0.348
```

$mathscor_{ij} = \beta_{0j} + \beta_{1j} * (hedres_{ij} - he\bar{d}res_j) + r_{ij}$

$\beta_{0j} = \gamma_{00} + u_{0j}$

$\beta_{1j} = \gamma_{10} + u_{1j}$

$\gamma_{00} = 49.8824212$ is the estimated grand mean math score.

$\gamma_{10} = 1.9372192$ is the average effect of hedres on math score.

$var(r_{ij}) = 40.391 = s^2$ is the within-group variance.

$var(u_{0j}) = 28.11$ is the between-group variance of the intercept β_{0j}.

$var(u_{1j}) = 9.1$ is the between-group variance of the slope β_{1j}.

Compared to Model 1, s^2 was reduced by 31%.

4. Based on your results for step 3, would it make sense to eliminate the random effect for the home-resources slope (u_{1j})? Why or why not? Justify your decision.

If we're eliminating the random effect then we're saying that the slope of the hedres effect has the same variance across all schools. The between-group variance of the slope was estimated to be 9.1, with a 3.017 variance. This would pass a standard significance test, so we should not eliminate the random effect.

5. MODEL 3: Finally, run a *contextual or conditional model*, and add the school academic press (**centered at the grand-mean**) as a level-2 predictor of both the level-1 intercepts and the home-resources slopes. Report and interpret all the parameters. **Compared to Model 2**, was the variability in the intercepts

between schools reduced with the addition of the academic press variable? What was the proportion reduction in this variance? What about the variability in home-resources slopes? Compared to Model 2, what proportion of this variance was accounted for by academic press?

```
acadpres_grand_mean <- data %>%
  summarize(acadpres.grand.mean = mean(acadpres))

data$acadpres.grand.mean.cen <- data$acadpres - acadpres_grand_mean$acadpres.grand.mean

model3 <- lmer(mathscor ~ hedres.group.mean.cen * acadpres.grand.mean.cen +
                 (hedres.group.mean.cen|schoolid), data = data)
model3.sum <- summary(model3)
model3.sum
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: mathscor ~ hedres.group.mean.cen * acadpres.grand.mean.cen +
##     (hedres.group.mean.cen | schoolid)
##    Data: data
##
## REML criterion at convergence: 1167.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.5489 -0.7331  0.1242  0.6339  2.0463
##
## Random effects:
##  Groups   Name                  Variance Std.Dev. Corr
##  schoolid (Intercept)            2.999   1.732
##           hedres.group.mean.cen  9.268   3.044    -0.25
##  Residual                       40.366   6.353
## Number of obs: 176, groups:  schoolid, 10
##
## Fixed effects:
##                                              Estimate Std. Error t value
## (Intercept)                                   49.7355     0.7285   68.27
## hedres.group.mean.cen                          1.9808     1.0660    1.86
## acadpres.grand.mean.cen                        3.9401     0.5979    6.59
## hedres.group.mean.cen:acadpres.grand.mean.cen  -0.7804     0.8993   -0.87
##
## Correlation of Fixed Effects:
##             (Intr) hdr... acd...
## hdrs.grp.m. -0.170
## acdprs.gr.. -0.013  0.005
## hdrs...:...  0.004  0.031 -0.163
```

$mathscor_{ij} = \beta_{0j} + \beta_{1j} * (hedres_{ij} - hed\bar{r}es_j) + r_{ij}$

$\beta_{0j} = \gamma_{00} + \gamma_{01} * (acadpres_j - aca\bar{d}pres) + u_{0j}$

$\beta_{1j} = \gamma_{10} + \gamma_{11} * (acadpres_j - aca\bar{d}pres) + u_{1j}$

$\gamma_{00} = 49.735459$ is the estimated grand mean math score.

$\gamma_{10} = 1.9808321$ is the average effect of hedres on math score.

$\gamma_{01} = 3.9401151$ is the average effect of acadpres on math score.

$\gamma_{11} = -0.7804058$ is the average effect of acadpres on the average effect of hedres on math score, i.e. the effect of acadpres on the slope.

$var(r_{ij}) = 40.366 = s^2$ is the within-group variance.

$var(u_{0j}) = 3.00$ is the between-group variance of the intercept β_{0j}.

$var(u_{1j}) = 9.27$ is the between-group variance of the slope β_{1j}.

Compared to Model 2, the variability in the intercepts between schools was reduced by 89%.

Compared to Model 2, the variability in home-resources slopes between schools increased by 2%.