Correlated Data

Mine Dogucu
Noah Johnson

April 5, 2018

General Linear Models require responses to be approximately normally distributed and independent. Through generalized linear models, we learned about handling responses that are not normally distributed (e.g. Poisson, Binomial). From now on we learn about multilevel models / linear mixed effect models / hierarchical linear models that can model response that violate the independence assumption. These models have responses that are correlated thus not independent.

```
hsb <- read.csv('hsb.csv')
str(hsb)
```

```
## 'data.frame':    7185 obs. of  10 variables:
##  $ schoolid: int  1224 1224 1224 1224 1224 1224 1224 1224 1224 1224 ...
##  $ minority: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ female  : int  1 1 0 0 0 0 1 0 1 0 ...
##  $ ses     : num  -1.528 -0.588 -0.528 -0.668 -0.158 ...
##  $ mathach : num  5.88 19.71 20.35 8.78 17.9 ...
##  $ size    : int  842 842 842 842 842 842 842 842 842 842 ...
##  $ sector  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ pracad  : num  0.35 0.35 0.35 0.35 0.35 0.35 0.35 0.35 0.35 0.35 ...
##  $ disclim : num  1.6 1.6 1.6 1.6 1.6 ...
##  $ himinty : int  0 0 0 0 0 0 0 0 0 0 ...
```

In this class and in the next few, we will be using the hsb.csv dataset. We will try to understand math achievement (mathach) of students based on their socio-economic status (ses). We will use the notation Yij for math achievement of ith student in the jth school. There are **7185** students within **160** schools in this dataset.

Calculate the average math achievement score for all students in the dataset.

```
avg_mathach <- hsb %>% summarise(avg_mathach = mean(mathach))
```

$Y_{..} = 12.7478526$

Pick a random school (make sure each school has equal probability to be selected). For **only** the school you selected, find the mean math achievement.

```
random_school_id <- hsb$schoolid %>%
  unique() %>%
  sample(1)

print(random_school_id)
```

```
## [1] 1308
```

```
my_school <- hsb %>% filter(schoolid == random_school_id)

mean_school_math_score <- my_school %>%
  summarise(mathach = mean(mathach))
```

$Y_{.j} = 16.2555$

For your school, fit a general linear model where math achievement is the response and ses is the predictor. Record the coefficients:

```r
my_school_model <- lm(mathach ~ ses, data = my_school)
summary(my_school_model)
```

```
##
## Call:
## lm(formula = mathach ~ ses, data = my_school)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.6898  -2.5708  -0.2306   5.4841   8.7483
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   16.189      2.118   7.642 4.67e-07 ***
## ses            0.126      3.003   0.042    0.967
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.281 on 18 degrees of freedom
## Multiple R-squared:  9.78e-05,   Adjusted R-squared:  -0.05545
## F-statistic: 0.001761 on 1 and 18 DF,  p-value: 0.967
```

$\hat{\beta_{0j}}$: 16.1889592 $\hat{\beta_{1j}}$: 0.1260242

For your school, fit a general linear model where math achievement is the response and grand mean centered ses is the predictor. Record the coefficients:

```r
hsb$ses.grandmean_c <- hsb$ses - mean(hsb$ses)
my_school <- hsb %>% filter(schoolid == random_school_id)

my_school_model.grand <- lm(mathach ~ ses.grandmean_c, data = my_school)
summary(my_school_model.grand)
```

```
##
## Call:
## lm(formula = mathach ~ ses.grandmean_c, data = my_school)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.6898  -2.5708  -0.2306   5.4841   8.7483
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       16.189      2.118   7.643 4.66e-07 ***
## ses.grandmean_c    0.126      3.003   0.042    0.967
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.281 on 18 degrees of freedom
## Multiple R-squared:  9.78e-05,   Adjusted R-squared:  -0.05545
## F-statistic: 0.001761 on 1 and 18 DF,  p-value: 0.967
```

$\hat{\beta_{0j}}$: 16.1889773 $\hat{\beta_{1j}}$: 0.1260242

For your school, fit a general linear model where group mean centered math achievement is the response and ses is the predictor. Record the coefficients:

```
school_means <- hsb %>%
  group_by(schoolid) %>%
  summarise(ses.groupmean = mean(ses))

hsb <- merge(hsb, school_means, by = "schoolid")

hsb$ses.groupmean_c <- hsb$ses - hsb$ses.groupmean

my_school_model.group <- lm(ses.grandmean_c ~ ses, data = my_school)
summary(my_school_model.group)
```

```
## Warning in summary.lm(my_school_model.group): essentially perfect fit:
## summary may be unreliable

##
## Call:
## lm(formula = ses.grandmean_c ~ ses, data = my_school)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.335e-16 -5.085e-17 -1.134e-17  2.119e-17  2.879e-16
##
## Coefficients:
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept) -1.434e-04  3.320e-17 -4.318e+12   <2e-16 ***
## ses          1.000e+00  4.707e-17  2.125e+16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.844e-17 on 18 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 4.514e+32 on 1 and 18 DF,  p-value: < 2.2e-16
```

$\hat{\beta_{0j}}$: $-1.4335421 \times 10^{-4}$ $\hat{\beta_{1j}}$: 1

We will use group-mean centering moving on.

Different effect types

Empty Model - One-Way Random-Effect ANOVA

```
install_load('lme4')
```

```
## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following object is masked from 'package:tidyr':
##
##     expand
```

```r
model.null <- lmer(mathach ~ 1 + (1|schoolid), data=hsb)
summary(model.null)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: mathach ~ 1 + (1 | schoolid)
##    Data: hsb
##
## REML criterion at convergence: 47116.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0631 -0.7539  0.0267  0.7606  2.7426
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  schoolid (Intercept)  8.614   2.935
##  Residual             39.148   6.257
## Number of obs: 7185, groups:  schoolid, 160
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  12.6370     0.2444   51.71
```

Model Notation:

Parameter Estimates:

ICC