Problem Set 6

Nicholas R. Gonzalez

Worked on with Tanner Bentley.

Problem 1

Hints

```r
#install.packages("mlmRev")
library(mlmRev)
```

```
Loading required package: lme4

Loading required package: Matrix
```

```r
df <- Hsb82

View(Hsb82)

df$school <- as.factor(df$school)

unique(df$sector)
```

```
[1] Public   Catholic
Levels: Public Catholic
```

```r
df$sector <- relevel(factor(df$sector), ref = "Public")

levels(df$sector)
```

```
[1] "Public"   "Catholic"
```

a

Regarding school type, attending a catholic school is associated with a 2.25 increase in math achievement, suggesting a positive effect for Catholic schools and math skills

Regarding minority status, being a racial minority is associated with a 3.11 lower math score compared to whites. This highlights a drastic gap in peformance in math between whites, and non whites.

According to our data, females score 1.42 points lower than male students, suggesting a gender gap, but not a drastic one.

Higher socio-economic status is associated with an increase in math scores, by 2.364 points.

In short, being from a catholic school, suggest higher math performance, as well as being a white, affluent male.

```r
model <- lm(mAch ~ sector + minrty + sx + ses, data = df)

library(modelsummary)
```

```
`modelsummary` 2.0.0 now uses `tinytable` as its default table-drawing
  backend. Learn more at: https://vincentarelbundock.github.io/tinytable/

Revert to `kableExtra` for one session:

  options(modelsummary_factory_default = 'kableExtra')
  options(modelsummary_factory_latex = 'kableExtra')
  options(modelsummary_factory_html = 'kableExtra')

Silence this message forever:

  config_modelsummary(startup_message = FALSE)
```

```r
modelsummary(model)
```

b

For our data, clustering the standard errors, or clustering in general is appropriate because each case has many students that go to one school. So observations are not independent.

	(1)
(Intercept)	13.242
	(0.134)
sectorCatholic	2.255
	(0.149)
minrtyYes	−3.112
	(0.170)
sxFemale	−1.422
	(0.146)
ses	2.364
	(0.099)
Num.Obs.	7185
R2	0.197
R2 Adj.	0.196
AIC	46 536.2
BIC	46 577.4
Log.Lik.	−23 262.081
F	440.111
RMSE	6.16

	(1)
(Intercept)	13.242
	(0.222)
sectorCatholic	2.255
	(0.275)
minrtyYes	−3.112
	(0.271)
sxFemale	−1.422
	(0.207)
ses	2.364
	(0.130)
Num.Obs.	7185
R2	0.197
R2 Adj.	0.196
AIC	46 536.2
BIC	46 577.4
RMSE	6.16
Std.Errors	by: school_yr

c

The coefficient results stayed the same, but my standard errors increased by about 100%. This is because without clustering the standard errors assume each student is independent, therefore clustering helps us adjust for intra-school correlaiton

```r
library(estimatr)

df$school_yr <- as.numeric(factor(df$school))

model_clustered <- lm_robust(mAch ~ sector + minrty + sx + ses,
                             data = df,
                             clusters = school_yr)

modelsummary(model_clustered)
```

d

All results are within a hundredth of each other, and the coefficients remain the same.

```
#install.packages('sandwich')
#install.packages('lmtest')
library(sandwich)
library(lmtest)
```

```
Loading required package: zoo


Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

```
model <- lm(mAch ~ sector + minrty + sx + ses, data = df)

summary(model)
```

```
Call:
lm(formula = mAch ~ sector + minrty + sx + ses, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-20.2286  -4.5076   0.2104   4.7472  17.8078

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    13.24158    0.13386  98.924   <2e-16 ***
sectorCatholic  2.25492    0.14906  15.127   <2e-16 ***
minrtyYes      -3.11239    0.17029 -18.277   <2e-16 ***
sxFemale       -1.42166    0.14608  -9.732   <2e-16 ***
ses             2.36392    0.09946  23.768   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 6.166 on 7180 degrees of freedom
Multiple R-squared:  0.1969,	Adjusted R-squared:  0.1965
F-statistic: 440.1 on 4 and 7180 DF,  p-value: < 2.2e-16
```

```r
vcov_clustered <- vcovCL(model, cluster = ~ school_yr)

coeftest(model, vcov = vcov_clustered)
```

```
t test of coefficients:

               Estimate Std. Error  t value  Pr(>|t|)    
(Intercept)    13.24158    0.22076  59.9805 < 2.2e-16 ***
sectorCatholic  2.25492    0.27278   8.2665 < 2.2e-16 ***
minrtyYes      -3.11239    0.26829 -11.6009 < 2.2e-16 ***
sxFemale       -1.42166    0.20541  -6.9210 4.873e-12 ***
ses             2.36392    0.12933  18.2786 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

e

After the bootstrap, my standard errors increased. This is likely because the bootstrap approach led to larger variance. It does so by re-sampling the data. This may mean that effects were not well estimated before. We have a decent size sample however, so I am not sure that bootstrapping is necessary.

```r
library(multiwayvcov)
library(lmtest)

model <- lm(mAch ~ sector + minrty + sx + ses, data = df)
summary(model)
```

```
Call:
lm(formula = mAch ~ sector + minrty + sx + ses, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-20.2286  -4.5076   0.2104   4.7472  17.8078 
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    13.24158    0.13386  98.924   <2e-16 ***
sectorCatholic  2.25492    0.14906  15.127   <2e-16 ***
minrtyYes      -3.11239    0.17029 -18.277   <2e-16 ***
sxFemale       -1.42166    0.14608  -9.732   <2e-16 ***
ses             2.36392    0.09946  23.768   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.166 on 7180 degrees of freedom
Multiple R-squared:  0.1969,    Adjusted R-squared:  0.1965
F-statistic: 440.1 on 4 and 7180 DF,  p-value: < 2.2e-16
```

```r
set.seed(123)

vcov_boot <- cluster.boot(model, cluster = df$school_yr, R = 1000, parallel = TRUE)

coeftest(model, vcov = vcov_boot)
```

```
t test of coefficients:

               Estimate Std. Error  t value  Pr(>|t|)
(Intercept)    13.24158    0.22576  58.6541 < 2.2e-16 ***
sectorCatholic  2.25492    0.27869   8.0911 6.887e-16 ***
minrtyYes      -3.11239    0.27618 -11.2693 < 2.2e-16 ***
sxFemale       -1.42166    0.20596  -6.9026 5.543e-12 ***
ses             2.36392    0.12670  18.6578 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Problem 2

a

```r
pset6data <- function(m, n){
  beta0 <- 0.3
  beta1 <- 0.8
```

```
  cluster.id <- sort(rep(1:m,n))
  d <- as.data.frame(cluster.id)
  d$X <- c()
  d$Y <- c()
  d$v <- c()
  d$epsilon <- c()
  for(i in 1:m){ # For each group
    v <- rnorm(1, 0, 0.5) # Group component in U
    mu <- rnorm(1, 0, 0.5) # Group component in X
    for(j in 1:n){ # For each observation in the group
      d$v[(i-1)*n+j] <- v
      d$X[(i-1)*n+j] <- rnorm(1, 0, 1) + mu
      d$epsilon[(i-1)*n+j] <- rnorm(1, 0, 0.5) # Individual component in U
      d$Y[(i-1)*n+j] <- beta0 + beta1*d$X[(i-1)*n+j] + d$v[(i-1)*n+j] + d$epsilon[(i-1)*n+j]
    }
  }
  return(d)
}
```

```
set.seed(123)

scenario_one <- pset6data(10, 500)

model_one <- lm(Y ~ X, data = scenario_one)

model_one_cluster <- lm_robust(Y ~ X, data = scenario_one, clusters = scenario_one$cluster.i

scenario_two <- pset6data(50, 100)

model_two <- lm(Y ~ X, data = scenario_two)

model_two_cluster <- lm_robust(Y ~ X, data = scenario_two, clusters = scenario_one$cluster.i

scenario_three <- pset6data(100, 50)

model_three <- lm(Y ~ X, data = scenario_three)

model_three_cluster <- lm_robust(Y ~ X, data = scenario_three, clusters = scenario_one$cluste

scenario_four <- pset6data(500, 10)

model_four <- lm(Y ~ X, data = scenario_four)
```

```r
model_four_cluster <- lm_robust(Y ~ X, data = scenario_four, clusters = scenario_one$cluster

# cant do model summary with lm_robust?

summary(model_one)
```

```
Call:
lm(formula = Y ~ X, data = scenario_one)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.57514 -0.43010 -0.00631  0.44748  2.09782 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.129175   0.009178   14.07   <2e-16 ***
X           0.686644   0.008535   80.45   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6404 on 4998 degrees of freedom
Multiple R-squared:  0.5643,	Adjusted R-squared:  0.5642 
F-statistic:  6472 on 1 and 4998 DF,  p-value: < 2.2e-16
```

```r
summary(model_one_cluster)
```

```
Call:
lm_robust(formula = Y ~ X, data = scenario_one, clusters = scenario_one$cluster.id)

Standard error type:  CR2 

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper    DF
(Intercept)   0.1292    0.13018  0.9923 3.473e-01  -0.1659   0.4243 8.879
X             0.6866    0.03106 22.1074 5.564e-09   0.6161   0.7572 8.749

Multiple R-squared:  0.5643 ,	Adjusted R-squared:  0.5642 
F-statistic: 488.7 on 1 and 9 DF,  p-value: 3.746e-09
```

```r
summary(model_two)
```

```
Call:
lm(formula = Y ~ X, data = scenario_two)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.34121 -0.48594  0.00005  0.48693  2.30147 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.412633   0.009815   42.04   <2e-16 ***
X           0.797122   0.008644   92.22   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6933 on 4998 degrees of freedom
Multiple R-squared:  0.6299,	Adjusted R-squared:  0.6298 
F-statistic:  8505 on 1 and 4998 DF,  p-value: < 2.2e-16
```

```r
summary(model_two_cluster)
```

```
Call:
lm_robust(formula = Y ~ X, data = scenario_two, clusters = scenario_one$cluster.id)

Standard error type:  CR2 

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper    DF
(Intercept)   0.4126    0.08604   4.796 9.807e-04   0.2180   0.6073 8.996
X             0.7971    0.03792  21.022 7.559e-09   0.7111   0.8831 8.832

Multiple R-squared:  0.6299 ,	Adjusted R-squared:  0.6298 
F-statistic: 441.9 on 1 and 9 DF,  p-value: 5.848e-09
```

```r
summary(model_three)
```

```
Call:
lm(formula = Y ~ X, data = scenario_three)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.15985 -0.48278  0.00116  0.46265  2.37445 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.358494   0.009999   35.85   <2e-16 ***
X           0.788750   0.008927   88.35   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7067 on 4998 degrees of freedom
Multiple R-squared:  0.6097,    Adjusted R-squared:  0.6096 
F-statistic:  7806 on 1 and 4998 DF,  p-value: < 2.2e-16
```

```r
summary(model_three_cluster)
```

```
Call:
lm_robust(formula = Y ~ X, data = scenario_three, clusters = scenario_one$cluster.id)

Standard error type:  CR2 

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper    DF
(Intercept)   0.3585    0.03816   9.394 6.011e-06   0.2722   0.4448 8.999
X             0.7887    0.01366  57.744 7.604e-13   0.7578   0.8197 8.970

Multiple R-squared:  0.6097 ,   Adjusted R-squared:  0.6096 
F-statistic:  3334 on 1 and 9 DF,  p-value: 7.055e-13
```

```r
summary(model_four)
```

```
Call:
lm(formula = Y ~ X, data = scenario_four)

Residuals:
```

```
     Min       1Q   Median       3Q      Max 
-2.18140 -0.46428  0.01783  0.46658  2.52354 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.285302   0.009895   28.83   <2e-16 ***
X           0.794368   0.008763   90.65   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6996 on 4998 degrees of freedom
Multiple R-squared:  0.6218,    Adjusted R-squared:  0.6217 
F-statistic:  8218 on 1 and 4998 DF,  p-value: < 2.2e-16
```

```r
summary(model_four_cluster)
```

```
Call:
lm_robust(formula = Y ~ X, data = scenario_four, clusters = scenario_one$cluster.id)

Standard error type:  CR2 

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper    DF
(Intercept)   0.2853    0.03782   7.544 3.525e-05   0.1998   0.3708 9.000
X             0.7944    0.01391  57.119 8.418e-13   0.7629   0.8258 8.968

Multiple R-squared:  0.6218 ,   Adjusted R-squared:  0.6217 
F-statistic:  3263 on 1 and 9 DF,  p-value: 7.78e-13
```

b

For our first scenario, the coefficients remain the same in both models, but the standard errors in the non-clustered model(0.0085) is smaller than in the clustered model (0.0311). This is as expected, from what we previously discussed about clustering.

This is also true for scenario two, where standard error in the non-clustered model (0.0086) is again smaller than the clustered model (0.0379).

This same logic upholds across every scenario, but with their respective results.

For scenarios three and four, again the coefficients are unchanged. However the ratio changes. The ratios are S1: 3.65, S2: 4.11, S3: 1.54, S4: 1. 58. This is because the degree of clustering, or the amount of grouping differs with each scenario. Also the ratio like changes based on the intra-cluster effects. Which it makes sense as scenarios three and four have more clusters, but less units.

Is it be safe to say that larger ratios mean stronger intra-cluster effects wheresas smaller eans weaker?