Stat 414 - Day 7
Correlated Observations

Last Time

- We often focus on fitting a linear model to predict the mean response depending on predictor variables. After adjusting for those "fixed effects," there is often a lot of information in the residuals, e.g., what patterns are still left unexplained. Perhaps there is an important variable missing from our model. Perhaps there are "variance covariates" that can explain variation in the variation of the residuals ("heterogeneity"). Perhaps we are most interested in adjusting the standard errors of the regression coefficients to improve the appropriateness of our p-values and confidence intervals.
- Multiple regression models work beautifully for both randomized experiments and observational studies. Randomized experiments are often designed to have "orthogonality" among the predictors so they explain distinct sources of variation in the response. With observational studies on the other hand, we often have to deal with the "overlap" of variation explain by the different predictors (e.g., multicollinearity, sequential vs. adjusted tests). In particular, we must always recognize that slope coefficients are "adjusted for" other variables in the model.
- Don't forget to consider visualizations (e.g., added-variable plots) as a tool for explaining your model.
- ICC can also be interpreted as the amount of correlation in pairs of observations within the same group. Keep in mind how you would manually calculate this number (e.g., find all possible pairs, how correlated are the two sets of responses) and that this is different from the correlation coefficient of two variables.
- Also keep in mind the distinction between the variance-covariance matrix of the parameter estimates (e.g., $V(\hat{\beta})$ and and the variance-covariance matrix of the residuals $V(\epsilon_i)$ which impacts the variance-covariance matrix of the responses $V(Y_i)$. Initially, we assumed $V(\epsilon_i) = \sigma^2 I$ and then we looked at ways to allow those diagonal elements to not all be the same (e.g., gls). Next we will focus on the off-diagonal elements.

Example 1: Finger Tapping Study

Caffeine is widely used as a stimulant – but are there other ways to get the same effects, with little to no downside? To begin to answer this question, a study compared the effects of caffeine with theobromine, which is the active chemical naturally found in chocolate and is an alkaloid with a similar molecular structure and effects on people as caffeine (Scott & Chen, 1944, "Comparison of Action of 1-Ethyl Theobromine and Caffeine in Animals and Man," *the Journal of Pharmacology and Experimental Therapeutics*). To measure the effects of these two different chemicals, the researchers trained subjects to tap their fingers in such a way that the rate could be measured. After learning/practicing this type of finger tapping, participants took either took a caffeine pill (200 mg), a theobromine pill (200 mg), or a placebo, and then their finger tapping rate was measured two hours later.

```
fingertapstudy = read.table("http://www.isi-stats.com/isi2/data/Fingertap.txt","\t"
, header=TRUE)
attach(fingertapstudy)  #this is an optional and sometimes looked-down-upon method
for letting R know which data file you are using so you don't have to use the data
```

```
file name every time
summary(fingertapstudy)
      Taps      Stimulant         participant
 Min.   :446   Length:12          Length:12
 1st Qu.:454   Class :character   Class :character
 Median :469   Mode  :character   Mode  :character
 Mean   :474
 3rd Qu.:485
 Max.   :523
var(Taps)
[1] 607.5
#We will use maximum likelihood estimation today (REML)
#install.packages("nlme")
library(nlme)
summary(model0 <- gls(Taps ~ 1))
Generalized least squares fit by REML
  Model: Taps ~ 1
  Data: NULL
    AIC BIC logLik
  108.2 109  -52.1

Coefficients:
            Value Std.Error t-value p-value
(Intercept)   474     7.115   66.62       0

Standardized residuals:
    Min      Q1     Med      Q3     Max
-1.1361 -0.7912 -0.2029  0.4362  1.9881

Residual standard error: 24.65
Degrees of freedom: 12 total; 11 residual
#variance-covariance matrix of the residuals
#install.packages("nlraa")
vcmatrix0 = nlraa::var_cov(model0); vcmatrix0[1:5, 1:5]
      [,1]  [,2]  [,3]  [,4]  [,5]
[1,] 607.5   0.0   0.0   0.0   0.0
[2,]   0.0 607.5   0.0   0.0   0.0
[3,]   0.0   0.0 607.5   0.0   0.0
[4,]   0.0   0.0   0.0 607.5   0.0
[5,]   0.0   0.0   0.0   0.0 607.5
```

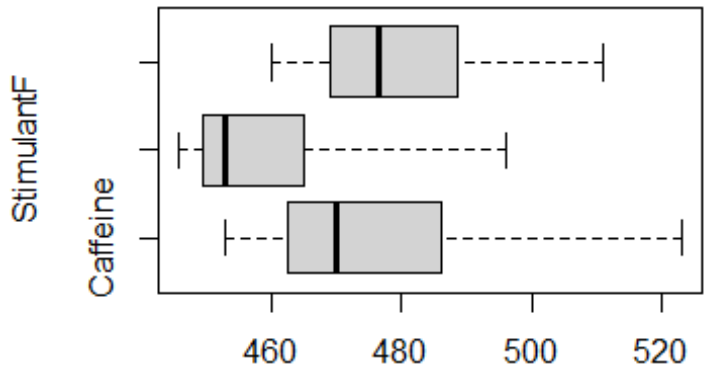
To make some math easier, I want to use effect-coding for the categorical variables today. We can change the coding from the get go:

```
# Set the contrast for the factor 'group'
participantF = as.factor(participant)
contrasts(participantF) <- "contr.sum"
contrasts(participantF)
  [,1] [,2] [,3]
A    1    0    0
B    0    1    0
```

```
C    0    0    1
D   -1   -1   -1
StimulantF = as.factor(Stimulant)
contrasts(StimulantF) <- "contr.sum"
contrasts(StimulantF)
            [,1] [,2]
Caffeine       1    0
Placebo        0    1
Theobromine   -1   -1
```

Consider a one-way ANOVA on the stimulants:

```
load(url("https://www.rossmanchance.com/iscam4/ISCAM.RData"))
iscamsummary(Taps, StimulantF)
            Missing n Min    Q1 Median    Q3 Max Mean    SD Skewness
Caffeine          0 4 453 462.8  470.0 486.2 523  479 30.58    0.878
Placebo           0 4 446 449.8  453.0 465.2 496  462 22.96    1.066
Theobromine       0 4 460 469.0  476.5 488.5 511  481 21.77    0.634
iscamboxplot(Taps, StimulantF)
```




```
summary(aov(Taps ~ StimulantF))
            Df Sum Sq Mean Sq F value Pr(>F)
StimulantF   2    872     436    0.68   0.53
Residuals    9   5810     646
summary(modelA <- gls(Taps ~ StimulantF))
Generalized least squares fit by REML
  Model: Taps ~ StimulantF
  Data: NULL
    AIC   BIC logLik
  98.13 98.92 -45.06

Coefficients:
            Value Std.Error t-value p-value
(Intercept)   474     7.335   64.63  0.0000
StimulantF1     5    10.373    0.48  0.6413
StimulantF2   -12    10.373   -1.16  0.2771
```

```
 Correlation:
            (Intr) StmlF1
StimulantF1  0.0
StimulantF2  0.0   -0.5

Standardized residuals:
    Min      Q1     Med      Q3     Max
-1.0233 -0.5412 -0.3149  0.2952  1.7318

Residual standard error: 25.41
Degrees of freedom: 12 total; 9 residual
vcmatrixa = nlraa::var_cov(modelA); vcmatrixa[1:5, 1:5]
      [,1]  [,2]  [,3]  [,4]  [,5]
[1,] 645.6   0.0   0.0   0.0   0.0
[2,]   0.0 645.6   0.0   0.0   0.0
[3,]   0.0   0.0 645.6   0.0   0.0
[4,]   0.0   0.0   0.0 645.6   0.0
[5,]   0.0   0.0   0.0   0.0 645.6
```

Formulas

With effect coding and equal group sizes ($N = Gn$), the standard deviation of the intercept is $SD(\bar{y}) = \sigma/\sqrt{N}$ and the standard deviation of the slope coefficients is $\sigma \times \sqrt{(G-1)/N}$ where G is the number of groups.

(a) What is the intercept? Why? What is its standard error? What about the treatment coefficients? Is the difference among the stimulants statistically significant? (Be clear how you are deciding.) The researcher insists there is a mistake because they know there is a difference among these treatments; how would you respond?

The intercept is ybar (balanced design, effect coding) with se 25.41/sqrt(12) and, with effect coding, the treatment coefficient SEs are 25.41*sqrt(G-1/N) = 10.373 where G = number of treatments and N is overall sample size. The difference between the stimulants is not statistically significant, because the remaining unexplained variability is still quite large. We have a small sample size in this study and therefore low power.

Consider a one-way ANOVA on the participants.

```
#using effect coding
summary(modelB <- gls(Taps ~ participantF))
Generalized least squares fit by REML
  Model: Taps ~ participantF
  Data: NULL
    AIC   BIC logLik
  79.98 80.38 -34.99

Coefficients:
              Value Std.Error t-value p-value
(Intercept)     474     3.541  133.84  0.0000
```

```
participantF1   -17     6.134   -2.77  0.0242
participantF2   -15     6.134   -2.45  0.0402
participantF3    -4     6.134   -0.65  0.5326

 Correlation:
             (Intr) prtcF1 prtcF2
participantF1  0.000
participantF2  0.000 -0.333
participantF3  0.000 -0.333 -0.333

Standardized residuals:
     Min       Q1      Med       Q3      Max
-1.22271 -0.71325  0.08151  0.65211  1.22271

Residual standard error: 12.27
Degrees of freedom: 12 total; 8 residual
vcmatrixb = nlraa::var_cov(modelB); vcmatrixb[1:5, 1:5]
      [,1]  [,2]  [,3]  [,4]  [,5]
[1,] 150.5   0.0   0.0   0.0   0.0
[2,]   0.0 150.5   0.0   0.0   0.0
[3,]   0.0   0.0 150.5   0.0   0.0
[4,]   0.0   0.0   0.0 150.5   0.0
[5,]   0.0   0.0   0.0   0.0 150.5
```

(b) What is the estimated intercept? Why? What is its standard error? Is there statistically significant person to person variation? Why might this be relevant to our real research question comparing the treatments?

with effect coding, intercept still estimate overall mean 12.268/sqrt(12) = 3.541. The p-value for testing the equality of all the participants' means = .0024 so yes, significant person to person variation in finger tapping rates. RCBD will allow a more direct comparison of the finger tap rates across treatments within each participant…

(c) Include the participant variable in the model. How should this change the ANOVA table?

will separate out the SS participants out of SSError but don't expect the coefficients of the treatments to change because of orthogonality in the RCBD (equal cell sizes).

```
summary(modelC <- gls(Taps ~ StimulantF + participantF, data = fingertapstudy))
Generalized least squares fit by REML
  Model: Taps ~ StimulantF + participantF
  Data: fingertapstudy
    AIC   BIC logLik
  66.15 64.69 -26.07

Coefficients:
             Value Std.Error t-value p-value
(Intercept)    474     2.147  220.74  0.0000
StimulantF1      5     3.037    1.65  0.1508
```

```
StimulantF2     -12     3.037   -3.95  0.0075
participantF1   -17     3.719   -4.57  0.0038
participantF2   -15     3.719   -4.03  0.0069
participantF3    -4     3.719   -1.08  0.3235

 Correlation:
              (Intr) StmlF1 StmlF2 prtcF1 prtcF2
StimulantF1    0.000
StimulantF2    0.000 -0.500
participantF1  0.000  0.000  0.000
participantF2  0.000  0.000  0.000 -0.333
participantF3  0.000  0.000  0.000 -0.333 -0.333

Standardized residuals:
       Min         Q1        Med         Q3        Max
-1.210e+00 -5.041e-01 -4.202e-14  5.377e-01  1.075e+00

Residual standard error: 7.439
Degrees of freedom: 12 total; 6 residual
anova(modelC)
Denom. DF: 6
             numDF F-value p-value
(Intercept)      1   48725  <.0001
StimulantF       2       8  0.0210
participantF     3      33  0.0004
vcmatrixc = nlraa::var_cov(modelC); vcmatrixc[1:5, 1:5]
      [,1]  [,2]  [,3]  [,4]  [,5]
[1,] 55.33  0.00  0.00  0.00  0.00
[2,]  0.00 55.33  0.00  0.00  0.00
[3,]  0.00  0.00 55.33  0.00  0.00
[4,]  0.00  0.00  0.00 55.33  0.00
[5,]  0.00  0.00  0.00  0.00 55.33
```

(d) Is the treatment variable now statistically significant? Why? How would you interpret the coefficient of placebo? What are the standard errors of the placebo and theobromine coefficients?

Yes, after adjusting for participants, p-value for Stimulant is .021 < .05. The standard errors of the coefficients are smaller. The coefficient of the placebo is how much lower the placebo treatment is from the overall mean on average (ybar_placebo vs. ybar). SE for a participant coefficient is 7.439 *and for treatment is 7.439* and for intercept is 7.439/.

We note that each subject participated in all 3 treatments, in random order, giving us a randomized block design.

(e) Is it reasonable to consider the observations in this study independent from each other? What should the variance-covariance matrix of the residuals look like?

We expect observations within the same person to be correlated, but uncorrelated with other individuals. This will give us a 'block diagonal' variance-covariance matrix.

Calculate and interpret the intraclass correlation for the subjects in the stimulant study.

```
#install.packages("ICC")
ICC::ICCbare(y = Taps, x = participantF)
[1] 0.7877
```

There is a strong correlation (.79) between the finger tapping rates with repeated observations on the same individual.

(f) Does this amount of correlation seem 'meaningful? How does this compare to the intraclass correlation coefficient we found with the Pace of Life study? Does that make sense in context? Explain.

This is larger than what we saw before. We expect repeat observations on the same individual to be more highly correlated that observations on different individuals living in the same city.

So let's use generalized least squares model to model the intraclass correlation of repeat observations on the same individual.

So we are now changing the assumptions of the basic regression model.

$$cor(\epsilon_{ij}, \epsilon_{i'j}) = \rho \neq 0$$

The following model assumes "compound symmetry" (equal variances, equal covariances).

Let's first just look at the participants variable.

```
#install.packages("nlme")
modelD <- nlme::gls(Taps ~ 1,
        corr = corCompSymm(form = ~1 | participantF))
summary(modelD)
Generalized least squares fit by REML
  Model: Taps ~ 1
  Data: NULL
    AIC   BIC logLik
  102.3 103.5 -48.17

Correlation Structure: Compound symmetry
 Formula: ~1 | participantF
 Parameter estimate(s):
   Rho
0.7877

Coefficients:
            Value Std.Error t-value p-value
(Intercept)   474     12.34   38.43       0

Standardized residuals:
    Min      Q1     Med      Q3     Max
-1.0516 -0.7323 -0.1878  0.4037  1.8402
```

```
Residual standard error: 26.63
Degrees of freedom: 12 total; 11 residual
anova(modelD)
Denom. DF: 11
            numDF F-value p-value
(Intercept)     1    1476  <.0001
#For comparison
#summary(model0 <- gls(Taps ~ 1))
#anova(model0, modelD)
```

(g) What is the estimated correlation coefficient? Does it look familiar? Is it statistically significant? What is the residual standard error? How has the standard error for the intercept changed (from model0)? Why? How would you explain this?

'Rho' is estimated to be 0.78772. The estimated intercept is still $\bar{y}$ with SE 12.33 The variance-covariance matrix has that block diagonal pattern. The residual standard error is 26.63. The standard error of the itnercept has increased because it now reflects the smaller 'effective sample size' from the correlated observations. $\sigma/\sqrt{N} \times \sqrt{1 + (1 - ICC) \times N}$. If we compare the model with and without 'rho' the 'full model' is significantly better from the likelihood ratio test (H_0: $\rho = 0$ vs. H_a: $\rho \neq 0$, df = 1, p-value = .0051).

What does our variance-covariance matrix look like?

Detour

- $Cor(X,Y) = Cov(X,Y)/[SD(X)SD(Y)]$

```
#Note, this doesn't work for the intercept only model
#vcmatrixd = nlraa::var_cov(modelD); vcmatrixd[1:5, 1:5]

#So we will use this instead
getVarCov(modelD, individual = 1)
Marginal variance covariance matrix
      [,1]  [,2]  [,3]
[1,] 709.0 558.5 558.5
[2,] 558.5 709.0 558.5
[3,] 558.5 558.5 709.0
  Standard Deviations: 26.63 26.63 26.63
```

(g2) What are we considering to be X and Y here? What are our estimates of X and Y? What is the estimated covariance?

X and Y represent two different observations on the same person so use $\hat{\sigma}$ = 26.63, so covariance = .7877 x 26.63 x 26.63 = 558.6

So let's add this correlation structure to our earlier model:

```
modelE <- nlme::gls(Taps ~ 1 + StimulantF + participantF,
          corr = corCompSymm(form = ~1 | participantF))
summary(modelE)
Generalized least squares fit by REML
  Model: Taps ~ 1 + StimulantF + participantF
```

```
  Data: NULL
    AIC   BIC logLik
  68.15 66.48 -26.07

Correlation Structure: Compound symmetry
 Formula: ~1 | participantF
 Parameter estimate(s):
Rho
  0

Coefficients:
              Value Std.Error t-value p-value
(Intercept)     474     2.147  220.74  0.0000
StimulantF1       5     3.037    1.65  0.1508
StimulantF2     -12     3.037   -3.95  0.0075
participantF1   -17     3.719   -4.57  0.0038
participantF2   -15     3.719   -4.03  0.0069
participantF3    -4     3.719   -1.08  0.3235

 Correlation:
              (Intr) StmlF1 StmlF2 prtcF1 prtcF2
StimulantF1    0.000
StimulantF2    0.000 -0.500
participantF1  0.000  0.000  0.000
participantF2  0.000  0.000  0.000 -0.333
participantF3  0.000  0.000  0.000 -0.333 -0.333

Standardized residuals:
       Min         Q1        Med         Q3        Max
-1.210e+00 -5.041e-01 -7.633e-15  5.377e-01  1.075e+00

Residual standard error: 7.439
Degrees of freedom: 12 total; 6 residual
```

(h) Has anything changed (vs. model C)? Why or why not?

This model is actually not that much different from when we just added Participant to the model. This is because the two options are doing essentially the same thing.

What about this model?

```
modelF <- nlme::gls(Taps ~ 1 + StimulantF ,
          corr = corCompSymm(form = ~1 | participantF))
summary(modelF)
Generalized least squares fit by REML
  Model: Taps ~ 1 + StimulantF
  Data: NULL
    AIC   BIC logLik
  88.51 89.49 -39.25

Correlation Structure: Compound symmetry
```

```
 Formula: ~1 | participantF
 Parameter estimate(s):
   Rho
0.9143

Coefficients:
            Value Std.Error t-value p-value
(Intercept)   474    12.336   38.43  0.0000
StimulantF1     5     3.037    1.65  0.1341
StimulantF2   -12     3.037   -3.95  0.0033

 Correlation:
            (Intr) StmlF1
StimulantF1  0.0
StimulantF2  0.0   -0.5

Standardized residuals:
    Min      Q1     Med      Q3     Max
-1.0233 -0.5412 -0.3149  0.2952  1.7318

Residual standard error: 25.41
Degrees of freedom: 12 total; 9 residual
```

(i) What do we learn?

once control for x, remaining errors are even more correlated, new se(ybar) 25.41/sqrt(12)*sqrt(1 + 2*.914). So this gives us the 'se inflation' from the correlated observations without specifying the difference in means across the participants. The previous model would be more different if we modelled a different covariance structure vs. 'group differences'

Note: Using the blocking variable both in the fixed effects part of the model and in the random effects part of the model is often essentially redundant, but hopefully you can see how it might help to do different things in the two places, e.g., using a more complicated covariance structure for the regions.

Example 2: Computer Problem 7 (due 7am Wednesday)

Recall our pace of life data

```
PaceData = read.table("https://www.rossmanchance.com/stat414/data/Pace.txt", header
=TRUE)
head(PaceData)
             City Heart Walk Talk Bank Watch    Region
1       Boston,MA    24   28   24   31    30 Northeast
2      Buffalo,NY    29   23   23   30    33 Northeast
3      NewYork,NY    31   24   18   29    32 Northeast
4 SaltLakeCity,UT    26   28   23   28    23      West
5     Columbus,OH    26   22   30   27    23   Midwest
6    Worcester,MA    20   25   24   26    27 Northeast
```

Let's center the walk variable

```
walk.c = PaceData$Walk - mean(PaceData$Walk)
summary(m1 <- lm(Heart ~ walk.c, data = PaceData))

Call:
lm(formula = Heart ~ walk.c, data = PaceData)

Residuals:
   Min     1Q Median     3Q    Max
-9.052 -3.283 -0.475  3.707 10.332

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   19.806      0.827   23.96   <2e-16 ***
walk.c         0.423      0.196    2.16    0.038 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.96 on 34 degrees of freedom
Multiple R-squared:  0.121, Adjusted R-squared:  0.095
F-statistic: 4.68 on 1 and 34 DF,  p-value: 0.0377
```

(a) How do we interpret the intercept? What is the SE of the intercept?

What is the ICC due to the regions?

```
ICC::ICCbare(y = PaceData$Heart, x = PaceData$Region)
[1] 0.1565
```

Allow the model to estimate a block-specific correlation.

```
summary(m2 <- gls(Heart ~ walk.c , data = PaceData,
        corr = corCompSymm(form = ~1 | Region)))
Generalized least squares fit by REML
  Model: Heart ~ walk.c
  Data: PaceData
    AIC   BIC logLik
  223.3 229.4 -107.6

Correlation Structure: Compound symmetry
 Formula: ~1 | Region
 Parameter estimate(s):
    Rho
0.05066

Coefficients:
             Value Std.Error t-value p-value
(Intercept) 19.806    0.9870  20.067  0.0000
walk.c       0.373    0.1976   1.887  0.0678

 Correlation:
       (Intr)
```

```
walk.c 0

Standardized residuals:
     Min       Q1      Med       Q3      Max
-1.80622 -0.68152 -0.07925  0.75459  2.04808

Residual standard error: 4.996
Degrees of freedom: 36 total; 34 residual
#install.packages("stargazer")
library(stargazer)
stargazer(m1, m2, type = "text")

=======================================================
                          Dependent variable:
                     ---------------------------------
                                  Heart
                            OLS           generalized
                                          least squares
                            (1)               (2)
-------------------------------------------------------
walk.c                    0.423**           0.373*
                          (0.196)           (0.198)

Constant                 19.810***         19.810***
                          (0.827)           (0.987)

-------------------------------------------------------
Observations                 36                36
R2                          0.121
Adjusted R2                 0.095
Log Likelihood                               -107.600
Akaike Inf. Crit.             223.300
Bayesian Inf. Crit.           229.400
Residual Std. Error   4.960 (df = 34)
F Statistic         4.675** (df = 1; 34)
=======================================================
Note:                       *p<0.1; **p<0.05; ***p<0.01
```

(b) What is the estimated correlation within regions? Why is it not the same as the ICC? How have the standard errors changed? Why do we expect SE(intercept) to increase?

Note: The standard error for the explanatory variable could go up or down depending on whether the explanatory variable varies mostly within or between regions. Here, the standard error decreased; it's like getting a more direct comparison of the treatment effects when there are on more homogeneous units. On the other hand, if most of the explanatory variable variation is between blocks, say the R^2 from regressing x on the Region was above say 0.70, we would expect to see SE inflation due to the design effect (it's like having too small of SD(X) within the blocks!)

Example 3: Back to finger tapping study

An alternative approach

The finger tapping study is a good example where we aren't really all that interested in the four participants themselves, we were just trying to control for that person-to-person variability, to help us assess the person-adjusted differences among the stimulants. In fact, we might be willing to consider the participants as a random sample?…

(a) Suppose we had a larger study with lots more participants. What would be a downside to including the participant variable in the model?

That would be a lot of coefficients or really a lot of 'degrees of freedom' to estimate all those different coefficients.

In a situation like this, one option is to treat person as a *random effect* rather than a *fixed effect*. This means we are going to treat these 4 participants not as (the only) 4 levels of a factor, but as a random sample from a population (if I did the study again, I would get 4 different participants). The assumption we are going to make is that the "participant effects" follow a normal distribution, centered at zero, with variance τ^2. Let's call these participant effects, u_j, so we have $u_j \sim N(0, \tau^2)$. Our model equation becomes: $Y_{ij} = \beta_0 + u_j + \epsilon_{ij}$ where $u_j \sim N(0, \tau^2)$ and $\epsilon_{ij} \sim N(0, \sigma^2)$. We also assume $cov(u_j, \epsilon_{ij}) = 0$.

Big deal, I changed β's to u's, but that is one way of saying we aren't considering the participant effects as parameters anymore. Instead, we replace them with one parameter, τ^2, which represents the participant-to-participant variation in the population of (potential) participants. This "small" change will have a large impact on the properties of the model.

(b) According to this model, what is $V(Y_{ij})$?

$$V(Y_{ij}) = \tau^2 + \sigma^2$$

(c) According to this model, what is $Cov(Y_{ij}, Y_{kj})$ (Two different observations in the same 'group'?

$$Cov(Y_{ij}, Y_{kj}) = Cov(u_j + \epsilon_{ij}, u_j + \epsilon_{kj}) = Cov(u_j, u_j) + Cov(\epsilon_{ij}, \epsilon_{kj}) = \tau^2$$

To fit this model, today we will use the "lme" command from the nlme package which you already have because it contains gls.

```
#library(nlme)
rm1 = lme(fixed = Taps ~ 1 , random = ~1 | participant, data = fingertapstudy, meth
od="REML")
#The notation (1|subject) is how we tell R to treat the participants as random effe
cts

summary(rm1)
```

```
Linear mixed-effects model fit by REML
  Data: fingertapstudy
    AIC   BIC logLik
  102.3 103.5 -48.17

Random effects:
 Formula: ~1 | participant
        (Intercept) Residual
StdDev:       23.63    12.27

Fixed effects:  Taps ~ 1
            Value Std.Error DF t-value p-value
(Intercept)   474     12.34  8   38.43       0

Standardized Within-Group Residuals:
    Min      Q1     Med      Q3     Max
-1.2496 -0.7895  0.1400  0.5698  1.3015

Number of Observations: 12
Number of Groups: 4
logLik(rm1)
'log Lik.' -48.17 (df=3)
```

(d) How many parameters are estimated in this model? How does the estimated intercept change? Standard error? What are the estimated variance components?

There 3 parameters: intercept, τ^2, σ^2. The estimated intercept is still $\bar{y}$ with SE = 12.335. $\hat{\tau}$ = 23.63 and $\hat{\sigma}$ = 12.27

We can view the estimated variance-covariance matrix for individual subjects.

```
getVarCov(rm1, subject = "1", type = "marginal")[[1]]
      1     2     3
1 709.0 558.5 558.5
2 558.5 709.0 558.5
3 558.5 558.5 709.0
```

And we can make R do the conversion to correlations

```
cov2cor(getVarCov(rm1, subject = "1", type = "marginal")[[1]])
       1      2      3
1 1.0000 0.7877 0.7877
2 0.7877 1.0000 0.7877
3 0.7877 0.7877 1.0000
```

So we have partitioned the total random variability into a variance component for the individual observations within each person (assumed to be the same across the participants) and a variance component for the participants. This also nicely induces a non-zero correlation between two observations from the same Level 2 units (this allows us to model dependence within the groups).

(e) Find the estimated "total variation" by summing $\hat{\tau}^2 + \hat{\sigma}^2$.

23.63^2 + 12.27^2 = 709 (the diagonal entries of the variance-covariance marix

(f) How much of this variation is due to the different participants?

23.63^2/(709) = 0.788, our ICC

Notes

Correlated data is encountered in nearly every field. In education, student scores from a particular teacher are typically more similar than scores of other students who have had a different teacher. Here we expect repeated measurements on the same individual to be more similar than finger tap measurements from other participants. In political polling, opinions from members of the same household are usually more similar than opinions of members from other randomly selected households. The intraclass correlation coefficient indicates how "reliably" we can predict an observation based on which group (e.g., subject) it comes from. If you have a larger intraclass correlation coefficient, the effective sample size is smaller.

- Used a "mixed effects" model allows us to account for the block differences as well as the correlation of observations within blocks/clusters simultaneously.
- Some packages/functions report the estimated variances, some the estimated standard deviations, some both.

Computer Problem 7 cont.

(c) Do a quick internet or ChatGPT search on "random vs. fixed effects." Submit a short summary of the distinction, AND bring with you to class on Wednesday. Be ready to justify whether you would treat Region as a fixed or random effect in the Pace of Life study.