Homework 2

BIOS6643 Fall 2021

Due Tues 10/5/2021 at midnight

Question 1. Principal Component Analysis

Consider the eNO data, and how we applied PCA to the data for graphical purposes (see Graphs slides). Determine the slope of the regression of Post (Y_2) on Pre (Y_1) values (i.e., a standard 'baseline as covariate' model), and compare this to the 'slope' of the $PC1$ axis. Compare the slopes numerically and superimpose the lines on a scatterplot of Post versus Pre values.

In order to do this, recall $PC1 = aY_1 + bY_2$, where a and b are chosen to maximize the variance of $PC1$ (recall $a = 0.51$, $b = 0.86$ for the data; see the slides).

Note: in terms of Y_2 versus Y_1, the 'slope' of the $PC1$ axis is simply b/a; to create a line to graph for $PC1$, you can have it go through the joint sample mean of Y_1 and Y_2. This exercise helps demonstrate the 'regression' principle in a regression line.

A few comments: First, in terms of the graph, $PC1$ is an axis rather than a line, just like Y_1 and Y_2. This is why we need to anchor it through something; it makes sense to have it go through the joint sample means of Y_1 and Y_2, just like the regression line does. This will allows us to determine an intercept for $PC1$ in addition to the slope, which we already know.

See the code below that walks through the calculations.

Note in the graph below I added the 95% confidence ellipse for the joint mean (like a confidence interval but generalizing to 2 dimensions). You only need to plot the 2 lines on the scatterplot for full credit (blue = PC1 'line', red = regression line). In this case there is not much 'regression' in the regression line.

Note that the slope of the regression line is $(SD_{post}/SD_{pre}) \times r$ and the slope of the $PC1$ line is SD_{post}/SD_{pre}; since r is close to 1, we do not see much difference between the two.

```
eno <- here::here("data", "eno_data.txt") %>%
  read.table(header = T,  sep = " ", skip = 0)

fit1 <-lm(eno_post ~ eno_pre, data = eno)
coef(fit1)
```

```
## (Intercept)     eno_pre
##   -8.229517    1.546124
```

```
## compute radius
N <- length(eno$eno_pre); n <- 2
f <- qf(0.95, n, N - n)
r <- sqrt((n * (N - 1) * f) / ((N - n) * N))

## covariance matrix
sigma <- mat.or.vec(2, 2)
sigma[1, 2] <- cov(eno$eno_pre, eno$eno_post); sigma[2, 1] <- sigma[1, 2]
sigma[1, 1] <- var(eno$eno_pre); sigma[2, 2] <- var(eno$eno_post)

## ellipse center (means)
mny1 <- mean(eno$eno_pre); mny2 <- mean(eno$eno_post)
## plot the data
matplot(eno$eno_pre, eno$eno_post,
  xlim = c(0, 180), ylim = c(0, 180),
  xlab = expression(mu[1] * " (eNO pre)"),
  ylab = expression(mu[2] * " (eNO post)"),
  pch = 1)

## add the ellipse
ellipse(center = c(mny1, mny2), shape = sigma, radius = r)

## indicate marginal sample means
segments(40, -10, 40, 53.7, lty = 2)
segments(-10, 53.7, 40, 53.7, lty = 2)

## Other Confidence ellipse info
eig <- eigen(sigma); corr <- cov2cor(sigma)

## Parts to answer the HW question
```

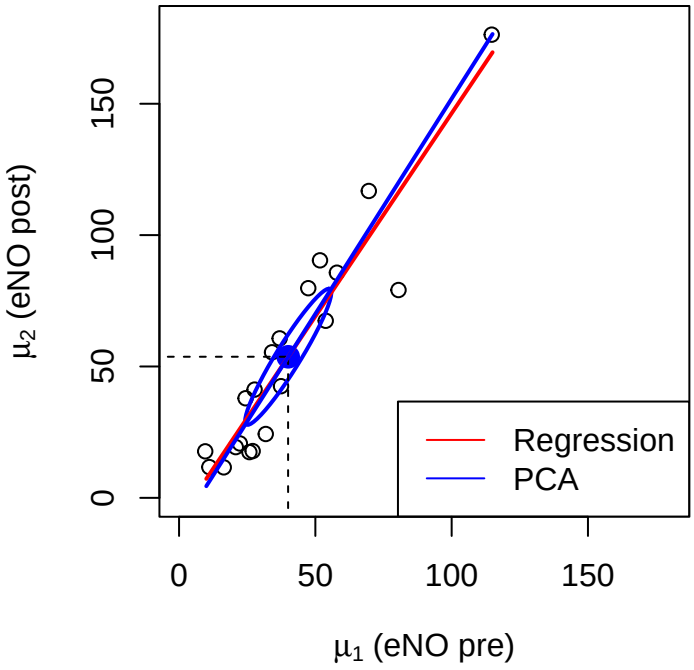
```
linreg <- lm(eno$eno_post ~ eno$eno_pre)
x <- c(10:115)
linregy <- coef(fit1)[1] + coef(fit1)[2] * x
lines(x, linregy, col = "red", lwd = 2)

## slope is the b term in the question
(slope <- sqrt(sigma[2, 2]) / sqrt(sigma[1, 1]))
```

```
## [1] 1.638786
```

```
## plug-in the b term to get intercept a
(yint <- mean(eno$eno_post) - mean(eno$eno_pre) * (slope))
```

```
## [1] -11.9402
```

```
pcy <- yint + slope * x
lines(x, pcy, col = "blue", lwd = 2)
legend("bottomright", lty = 1,
       col = c("red", "blue"),
       legend =c("Regression", "PCA"))
```



Question 2. GLM, GzLM, LMM, and likelihood functions, and Variance in LMM

a. In a paragraph, explain the difference between a general linear model (GLM; not a generalized linear model, which I denote with GzLM and which will be discussed more later) and a linear mixed model (LMM).

Basically, a general linear model (GLM) is for independent (e.g., cross-sectional or one-way ANOVA) data, and a linear mixed model (LMM) accounts for correlated data.

When there are violation on certain assumptions, such as independence or equal-variance assumption, GLM is not reasonable to be used directly; LMM is a powerful tool, allowing us to include more sophisticated terms: random effect *pmbb* and error $\boldsymbol{R}$ matrices. The GLM is a special case of the LMM when there are no random effects and the error covariance matrix is simple ($\sigma^2\boldsymbol{I}$).

Both modeling approaches are regression-type models, where we are trying to understand the relationship between an outcome and several. For the LMM, modeling the correlation (and covariance parameters in general) is usually a nuisance process (something we need to do but are not directly interested in). However, there are situations where we may be interested in random-effect estimates themselves, or even the other covariance parameter estimates.

b. In a short paragraph, explain the difference between a profiled likelihood and a restricted likelihood for a linear mixed model, and how and why they are used. Which one is a re-expression of the standard likelihood?

The common profiled likelihood for a linear mixed model is expressed completely in terms of the covariance parameters. This is accomplished by maximizing the likelihood conditioned on the covariance parameters, and then solving for the fixed effects. This leads to an algebraic form for $\hat{\boldsymbol{\beta}}$, expressed as a function of the covariance parameters. This form can then be substituted back in for $\boldsymbol{\beta}$, so that the likelihood is completely expressed in terms of covariance parameters, but it is intrinsically the same likelihood.

The restricted likelihood considers a linear form of the original $\boldsymbol{Y}$ that eliminates the fixed effects completely, so it is a different likelihood. The purpose is to get unbiased (or at least less biased) estimators of covariance parameters. The difficulty is there is no true mechanism to estimate the fixed effect parameters with the restricted likelihood, so what is typically done is that the ML algebraic form for $\hat{\boldsymbol{\beta}}$ is employed.

A profiled likelihood is a re-expression of the standard likelihood.

c. Derive $Var[\hat{\boldsymbol{\beta}}]$ in a full-rank linear mixed model, given the algebraic form of $\hat{\boldsymbol{\beta}}$ that is obtained via ML estimation.

NOTE: there are two types of variance, model-based and empirical (or sandwich estimator). The difference is whether the middle $\boldsymbol{V}$ is determined via the model or using squared residual quantities. To answer question c., work with the 'complete data' form of $\hat{\boldsymbol{\beta}}$.

The ML estimator has form $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^-\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{Y}$, which is a linear form of $\boldsymbol{Y}$. Since we are dealing with a model with full rank $\boldsymbol{X}$, then $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{Y}$. The linear form result says $Var[\boldsymbol{AY}] = \boldsymbol{A}Var[\boldsymbol{Y}]\boldsymbol{A}^t$; so let $\boldsymbol{A} = (\boldsymbol{X}^t\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{V}^{-1}$ and

$$
\begin{aligned}
Var(\hat{\boldsymbol{\beta}}) &= Var((\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{V}^{-1}\mathbf{Y}) && \text{ML estimate for } \boldsymbol{\beta}\\
&= [(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{V}^{-1}]Var(\mathbf{Y})[(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{V}^{-1}]^t && Var(\mathbf{AX}) = \mathbf{A}Var(\mathbf{X})\mathbf{A}^t\\
&= [(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{V}^{-1}]Var(\mathbf{Y})[(\mathbf{V}^{-1})^t(\mathbf{X}^t)^t((\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X})^{-1})^t] && (\mathbf{AB})^t = \mathbf{B}^t\mathbf{A}^t\\
&= [(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{V}^{-1}]Var(\mathbf{Y})[\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X})^{-1}] && \mathbf{A}^t = \mathbf{A} \text{ symmetric}\\
&= (\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}^t\mathbf{V}^{-1}\cancel{\mathbf{V}\mathbf{V}^{-1}}\mathbf{X})(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X})^{-1} && Var(\mathbf{Y}) = \mathbf{V}\\
&= (\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X})^{-1}\cancel{(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X})(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X})^{-1}} && \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}\\
&= (\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X})^{-1}
\end{aligned}
$$

Question 3. Models for Beta Carotene data

For the Beta Carotene data (see the description of the data and the data itself in another link in the Data module). For parts **a** and **b**, model *time* and *group* as class variables, and include $group \times time$. In order to account for repeated measures over *time*, specify the UN error covariance structure.

a. Conduct a test to compare the 30 and 60mg BASF trends over *time* to see if they differ, i.e., an interaction test, but only involving these 2 *groups*.

```
proc import DATAFILE='C:/Users/Goodgolden5/Desktop/beta_carotene_univar.csv'
DBMS=csv out=beta replace; run;

/*Interaction between time and BASF group*/
/*Using containment approach for degrees of freedom*/
proc mixed data= beta;
class prepar time;
model y= prepar time prepar*time /solution;
repeated / subject=ID(prepar) type=un;

contrast 'Interaction BASF 30mg and BASF 60mg'
/*terms     g1t0   g1t6   g1t8   g1t10  g1t12   g2t0   g2t6   g2t8   g2t10  g2t12   g3t0  g3t6  g3t8  g3t10  g3t12   g4t0   g4t6   g4t8  g4t10  g4t12 */
prepar*time 0      0      0      0      0       0      0      0      0      0       1     -1     0     0      0      -1     1      0     0      0,
prepar*time 0      0      0      0      0       0      0      0      0      0       1     0     -1     0      0      -1     0      1     0      0,
prepar*time 0      0      0      0      0       0      0      0      0      0       1     0      0    -1      0      -1     0      0     1      0,
prepar*time 0      0      0      0      0       0      0      0      0      0       1     0      0     0     -1      -1     0      0     0      1;
ods select contrasts;
   ods trace on;
   ods show;

run;

##                            The Mixed Procedure
##
##                                Contrasts
##
##                                          Num     Den
##  Label                                    DF      DF    F Value    Pr > F
##
##  Interaction BASF 30mg and BASF 60mg       4      19       3.34    0.0313
```

The results show that the overall interaction (involving all groups) is marginally significant at the 0.05 level, and the interaction involving only the BASF groups is slightly more significant. Note that the DDF methods here is 'between-within', the default in SAS when a repeated but not random statement is included. You could also use "subject=id" in this case and get the same DDF; "satterth" also yields the same DDF in this case.

Here we provide an "almost" equvalent model fitting with R. We will see some part of the statistics consistent with the SAS outputs. However the model fitting and parameter setup may cause inconsistencies. This is just an outline for a general strategy. Moreover you should explain your results accordingly with your outcomes.

```r
## import data beta
bc <- here::here("data", "beta_carotene_univar.csv") %>%
  read.csv() %>%
  janitor::clean_names() %>%
  mutate(time = as.integer(time))

## set up the control for convergence
## we included so many parameters
ctrl <- lmeControl(niterEM = 1000,
                   # opt="optim",
                   msMaxIter = 1000)

## model includes the random intercept and UN for residual
# mod1 <- lme(y ~ 1 + factor(time) * factor(prepar),
#             ## random intercept
#             random = ~1|id,
#             ## UN for correlation! no covariance
#             correlation = corSymm(form = ~1|id),
#             ## for unequal variance over time
#             weights = varIdent(form = ~1|time),
#             ## convergence setting
#             control = ctrl,
#             data = bc)

mod1 <- gls(y ~ 1 + factor(time) * factor(prepar),
            ## UN for correlation! not covariance
            correlation = corSymm(form = ~1|id),
            ## for unequal variance over time
            weights = varIdent(form = ~1|time),
            ## convergence setting
            control = ctrl,
            method = "REML",
            data = bc)
```

Please read the Rmd file for extra coding and information. Extra explanations and model fitting setup included in the Rmd file. We hope you can read and run the code by yourself.

We introduce three methods to get the contrast test.

1. Manually setup

```r
## fixed effects
# beta_hat <- fixed.effects(mod1)
beta_hat <- coef(mod1)
## fixed effects variance-covariance
C <- vcov(mod1) %>% round(digits = 2)

## check the Rmd file for a better version
##             [1]  [2] [3] [4]  [5]  [6] [7] [8] [9]   [10]  [11]   [12]   [13]  [14]  [15]   [16]   [17]  [18]  [19]   [20]
## coefs       Int  t6  t8  t10  t12  p2  p3  p4  t6:p2 t8:p2 t10:p2 t12:p2 t6:p3 t8:p3 t10:p3 t12:p3 t6:p4 t8:p4 t10:p4 t12:p4
p3_p4_t0 <- c( 0,   0,  0,  0,   0,   0,  1,  -1,  0,    0,    0,     0,     0,    0,    0,     0,     0,    0,    0,     0)
## p3_t6 <- c( 1,   1,  0,  0,   0,   0,  1,  0,   0,    0,    0,     1,     0,    0,    0,     0,     0,    0,    0,     0)
## p4_t6 <- c( 1,   1,  0,  0,   0,   0,  0,  1,   0,    0,    0,     0,     0,    0,    0,     0,     1,    0,    0,     0)
## p3_p4_t6 <- p3_t6 - p4_t6, similarly for other time points
p3_p4_t6 <- c( 0,   0,  0,  0,   0,   0,  1,  -1,  0,    0,    0,     0,     1,    0,    0,     0,     -1,   0,    0,     0)
p3_p4_t8 <- c( 0,   0,  0,  0,   0,   0,  1,  -1,  0,    0,    0,     0,     0,    1,    0,     0,     0,    -1,   0,     0)
p3_p4_t10 <- c(0,   0,  0,  0,   0,   0,  1,  -1,  0,    0,    0,     0,     0,    0,    1,     0,     0,    0,    -1,    0)
p3_p4_t12 <- c(0,   0,  0,  0,   0,   0,  1,  -1,  0,    0,    0,     0,     0,    0,    0,     1,     0,    0,    0,    -1)

contr1 <- cbind(p3_p4_t0, p3_p4_t6, p3_p4_t8, p3_p4_t10, p3_p4_t12)
rownames(contr1) <- rownames(C)

## build the contrast for the contrast
p34_t6_0 <- p3_p4_t6 - p3_p4_t0
p34_t8_0 <- p3_p4_t8 - p3_p4_t0
p34_t10_0 <- p3_p4_t10 - p3_p4_t0
p34_t12_0 <- p3_p4_t12 - p3_p4_t0

contr0 <- cbind(p34_t6_0, p34_t8_0, p34_t10_0, p34_t12_0)
## contrast point estimates to be
ce0 <- t(contr0) %*% beta_hat
## contrast variance covariance matrix
cov0 <- t(contr0) %*% C %*% contr0

## with both point estimates and standard deviation
## an anova or pairwise comparison can be performed
W0 <- t(ce0) %*% solve(cov0) %*% ce0
pchisq(W0, df = 4, lower.tail = FALSE)
```

```
          [,1]
[1,] 0.00965162
```

2. ‘multcomp::glht()‘

```r
## emmeans is a package cover
test1 <- multcomp::glht(mod1, t(contr0))
summary(test1, test = Chisqtest())
```

```
     General Linear Hypotheses

Linear Hypotheses:
             Estimate
p34_t6_0 == 0    -18.07
p34_t8_0 == 0    -51.87
p34_t10_0 == 0    22.20
p34_t12_0 == 0    48.80

Global Test:
  Chisq DF Pr(>Chisq)
1 13.36  4   0.009651
```

b. Conduct a test to compare to see if the 12 week - baseline value differs between the 4 *groups*.

```
proc import DATAFILE='C:/Users/Goodgolden5/Desktop/beta_carotene_univar.csv'
  replace out=beta dbms=csv; run;

proc mixed data= beta;
class prepar time;
model y= prepar time prepar*time /solution;
repeated / subject=ID(prepar) type=un;

contrast 'Interaction between Group and (12 weeks - baseline)'
/*terms     g1t0   g1t6   g1t8   g1t10  g1t12   g2t0   g2t6   g2t8   g2t10  g2t12   g3t0  g3t6  g3t8  g3t10  g3t12   g4t0   g4t6   g4t8  g4t10  g4t12 */
prepar*time -1     0      0      0      1       1      0      0      0      -1      0     0     0      0      0      0      0      0     0       0,
prepar*time -1     0      0      0      1       0      0      0      0      0       1     0     0      0      -1     0      0      0     0       0,
prepar*time -1     0      0      0      1       0      0      0      0      0       0     0     0      0      0      1      0      0     0       -1;
ods select contrasts;
   ods trace on;
   ods show;
run;
```

```
                          The Mixed Procedure

                              Contrasts

                                                      Num    Den
Label                                                  DF     DF   F Value

Interaction between Group and (12 weeks - baseline)     3     19      2.40

                              Contrasts

      Label                                            Pr > F

      Interaction between Group and (12 weeks - baseline)   0.0997
```

```
##             [1]  [2] [3] [4]  [5]  [6] [7] [8] [9]   [10]  [11]   [12]   [13]  [14]  [15]   [16]   [17]  [18]  [19]   [20]
## coefs       Int  t6  t8  t10  t12  p2  p3  p4  t6:p2 t8:p2 t10:p2 t12:p2 t6:p3 t8:p3 t10:p3 t12:p3 t6:p4 t8:p4 t10:p4 t12:p4
# p1_t12 <- c( 1,   0,  0,  0,   1,   0,  0,  0,  0,    0,    0,     0,     0,    0,    0,     0,     0,    0,    0,     0)
# p1_t0  <- c( 1,   0,  0,  0,   0,   0,  0,  0,  0,    0,    0,     0,     0,    0,    0,     0,     0,    0,    0,     0)
p1_t12_0 <- c( 0,   0,  0,  0,   1,   0,  0,  0,  0,    0,    0,     0,     0,    0,    0,     0,     0,    0,    0,     0)
p2_t12_0 <- c( 0,   0,  0,  0,   1,   0,  0,  0,  0,    0,    0,     1,     0,    0,    0,     0,     0,    0,    0,     0)
p3_t12_0 <- c( 0,   0,  0,  0,   1,   0,  0,  0,  0,    0,    0,     0,     0,    0,    0,     1,     0,    0,    0,     0)
p4_t12_0 <- c( 0,   0,  0,  0,   1,   0,  0,  0,  0,    0,    0,     0,     0,    0,    0,     0,     0,    0,    0,     1)

contr2 <- cbind(p2_t12_0 - p1_t12_0,
                p3_t12_0 - p1_t12_0,
                p4_t12_0 - p1_t12_0)
## contrast point estimates to be
(ce2 <- t(contr2) %*% beta_hat)
```

```
          [,1]
[1,] -58.00000
[2,]  72.46667
[3,]  23.66667
```

```
## contrast variance covariance matrix
cov2 <- t(contr2) %*% C %*% contr2

## with both point estimates and standard deviation
## an anova or pairwise comparison can be performed
W2 <- t(ce2) %*% solve(cov2) %*% ce2
pchisq(W2, df = 3, lower.tail = FALSE)
```

```
           [,1]
[1,] 0.06575128
```

```
## emmeans is a package cover
test2 <- multcomp::glht(mod1, t(contr2))
summary(test2, test = Chisqtest())
```

```
     General Linear Hypotheses

Linear Hypotheses:
       Estimate
1 == 0   -58.00
2 == 0    72.47
3 == 0    23.67

Global Test:
  Chisq DF Pr(>Chisq)
1 7.201  3    0.06575
```

c. Consider the model that uses *time* as continuous, with up to cubic effects, plus interactions between group and time (up to cubic). How does this model compare with the one that uses *time* as class (plus interactions)? Discuss in a paragraph.

```
mod2 <- gls(y ~ 1 + I(time) * factor(prepar) +
              I(time^2) * factor(prepar) +
              I(time^3) * factor(prepar) +
              I(time^4) * factor(prepar),
            correlation = corSymm(form = ~1|id),
            weights = varIdent(form = ~1|time),
            control = ctrl,
            method = "REML",
            data = bc)

mod1_ml <- update(mod1, method = "ML")
mod2_ml <- update(mod2, method = "ML")

## do not suppose to use LRT
## just to show that those two are
## the same models...
## there was a counting accident.
anova(mod1, mod2)
```

```
Warning in nlme::anova.lme(object = mod1, mod2): fitted objects with different
fixed effects. REML comparisons are not meaningful.

     Model df      AIC      BIC    logLik
mod1     1 35 1094.003 1183.389 -512.0017
mod2     2 35 1216.423 1305.809 -573.2116
AIC(mod1, mod2)
```

```
     df      AIC
mod1 35 1094.003
mod2 35 1216.423
```

The advantage of using time as a class variable is that each group by time interaction gets its own unique estimate. This means that there are not trend restrictions on the model; we are not constraining the model to a straight line or any other pattern. This is the most flexible model. Additionally, orthogonal contrasts allow for correct coefficients even when time is not equally spaced. The test is invariant to scale changes to the coefficients.

Yet, there are cases when a continuous time trend fits the model well. If imposing a linear, quadratic, cubic, or other time trend fits the data, then this simpler model may be sufficient. For example, if your results suggest an equally spaced linear time trend, then a simple linear time trend may be sufficient. Continuous time allows us to assess trends across small units of change in time. We can interpolate, allowing us to estimate short-term average changes.

d. Modeling the data using $Time0$ as a covariate value, with the remaining *times* as repeated measures on the outcome (6, 8, 10, 12 weeks). What are pros and cons of this approach, relative to using all measures as outcome values in a longitudinal model? In particular, focuses on the modeling of the repeated measures, how fixed effects need to be specified, and impact of modeling of *time* as class versus continuous.

```
bc2 <- bc %>%
  ## longer to wider
  pivot_wider(names_from = time,
              values_from = y) %>%
  ## wider to longer
  pivot_longer(cols = 4:7,
               names_to = "time",
               values_to = "y") %>%
  rename("baseline" = "0") %>%
  mutate(time = as.integer(time))


# mod3 <- lme(y ~ baseline + factor(time) * factor(prepar),
#             ## random intercept
#             random = ~1|id,
#             ## UN for correlation! no covariance
#             correlation = corSymm(form = ~1|id),
#             ## for unequal variance over time
#             weights = varIdent(form = ~1|time),
#             ## convergence setting
#             control = ctrl,
#             data = bc2)

mod4 <- gls(y ~ baseline + factor(time),
            ## UN for correlation! no covariance
            correlation = corSymm(form = ~1|id),
            ## for unequal variance over time
            weights = varIdent(form = ~1|time),
            ## convergence setting
            control = ctrl,
            data = bc2)
```

One advantage of using the baseline as a covariate is that you now have 4 equally spaced time points and you can use a simpler covariance structure, like the AR(1), which was built for equally spaced measures. With

this approach we can also establish a slope relationship between the outcome and baseline value. Using all 5 measures in a longitudinal model would allow you to estimate for times between 0 and 6 weeks using a smooth function, using polynomials and time as continuous. It gives us a fuller picture of changes over time, from 0 all the way up through 12 weeks (also see part e below).

e. For the model in part **d**, estimate the linear, quadratic and cubic trends for the orthogonal polynomial model that uses *time* as a class variable.

```
proc import DATAFILE='C:/Users/Goodgolden5/Desktop/beta_carotene_univar.csv'
  replace out=beta dbms=csv; run;

/*reshape data*/
proc sort data=beta; by prepar id time; run;
data bl; set beta; if time=0;
keep id prepar time y; rename y=y_bl;

proc sort data=bl; by id time;
proc sort data=beta; by id time;
data bigger; merge beta bl; by id;
if time^=0; run;

proc mixed data= bigger;
class prepar time;
model y= y_bl prepar time /solution;
repeated / subject=ID type=unr rcorr;
estimate 'linear' time -3 -1 1 3;
estimate 'quadratic' time 1 -1 -1 1;
estimate 'cubic' time -1 3 -3 1;
contrast 'Linear, Quadratic, and Cubic'
  time -3 -1 1 3,
  time 1 -1 -1 1,
  time -1 3 -3 1;
ods select estimates;
ods select contrasts;
   ods trace on;
   ods show;
run;
```

The Mixed Procedure

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
linear	26.2609	39.6786	18	0.66	0.5165
quadratic	15.6522	13.0081	18	1.20	0.2445
cubic	4.6957	32.3942	18	0.14	0.8864

Contrasts

Label	Num DF	Den DF	F Value	Pr > F
Linear, Quadratic, and Cubic	3	18	0.83	0.4956

So since I did not say otherwise, we can consider estimate these trends in the main effect for time, i.e., averaging over group. None of these effects are significant, which is not a surprise since the main effect of time itself was not significant. Importantly, though, if you perform the polynomial contrasts using the model that uses time 0 as an outcome, you will get different results, because the plasma levels increased in the subjects between 0 and 6 weeks; further increases were not significant, as demonstrated by the linear trend estimate.

Note: if you don't say anything then the default DDF method is "Between-Within" when there is a REPEATED statement and no RANDOM statement. You could specify something else; the Satterthwaite option is decent and a little more conservative, but yields pretty similar results. Again, results will be the same if you use "id" instead of "id(prepar)" as the subject in the REPEATED statement. The main reason I include it is in case we change to a RANDOM statement, for which there is a difference (and "id(prepar)" yields more intuitive DDF results via the default Containment method). The last contrast I just added on; this would be testing whether any of the polynomial trends are significant. You could even use 'Contrasts' to do the individual tests; results will be pretty similar (except you obviously won't get point estimates using that approach).

```
emm4_poly <- emmeans(mod4, ~factor(time))
```

```
Analytical Satterthwaite method not available; using appx-satterthwaite
```

```
contrast(emm4_poly, 'poly')
```

```
 contrast  estimate   SE   df t.ratio p.value
 linear        26.3 39.7 22.4   0.662  0.5148
 quadratic     15.7 13.0 22.6   1.203  0.2413
 cubic          4.7 32.4 21.6   0.145  0.8861

Degrees-of-freedom method: appx-satterthwaite
```

Question 4. Constrasts

Consider a study where *subjects* in 3 *groups* (e.g., race or treatment) are observed over 3 equally spaced *times* and some health outcome, y, is measured. Unless otherwise mentioned, include a random intercept for subjects to account for the repeated measures. For simplicity, use 2 *subjects* per *group*.

a. Consider modeling *group* and *time* as class variables, plus interaction. Write statistical models and the $\boldsymbol{X}$ matrix for the following cases.

b. No restriction placed on the model. i.e., write the less-than-full-rank statistical model.

$$Y_{grp=g,sub=i,time=t} = \mu_0 + \alpha_g + \tau_t + \gamma_{g\times t} + b_i + \epsilon_{g,i,t}$$
$$b_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_b^2)$$
$$\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$$
$$b_i \perp \epsilon_{ij}$$

```
## setup dataset
group <- rep(c("A", "B", "C"), each = 6)
time <- rep(c("1", "2", "3"))
id <- rep(1:6, each = 3)
## outcome y is just a placeholder
y <- "NA"
data_s <- cbind(id, group, time, y) %>% as.data.frame()

## formula and model.frame are important object
## for lme and gls model fitting.
form1 <- formula(y ~ I(group == "A") + I(group == "B") + I(group == "C") +
                   I(time == "1") + I(time == "2") + I(time == "3") +
                   group:time)
mod_f1 <- model.frame(form1, # Formula
                      # Data frame
                      data = data_s,
                      # Identifier of data records
                      SubjectId = id)
## calling design matrix
Xmtx1 <- model.matrix(form1, mod_f1)
colnames(Xmtx1) <- NULL; kable(Xmtx1, "simple")
```

1	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0
1	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0
1	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0
1	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0
1	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0
1	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0
1	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0
1	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0
1	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0
1	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0
1	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0
1	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0
1	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0
1	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0
1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1
1	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0
1	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0
1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1

ii. A set-to-0 restriction is placed on the parameters associated with highest levels.

$$Y_{ij} = \beta_0 + \beta_1 G_1 + \beta_2 G_2 + \beta_3 t_1 + \beta_4 t_2 + \beta_{13} G_1 t_1 + \beta_{14} G_1 t_2 + \beta_{23} G_2 t_1 + \beta_{24} G_2 t_2 + b_i + \epsilon_{ij}$$
$$b_i \overset{iid}{\sim} N(0, \sigma_b^2)$$
$$\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$
$$b_i \perp \epsilon_{ij}$$

```
form2 <- formula(y ~ 1 + group + time + group:time)
mod_f2 <- model.frame(form2, # Formula
                      # Data frame
                      data = data_s,
                      # Identifier of data records
                      SubjectId = id)
Xmtx2 <- model.matrix(form2, mod_f2,
                      contrasts.arg = list(group = "contr.SAS",
                      ## SAS uses the highest as reference group
                                           time = "contr.SAS"))
colnames(Xmtx2) <- NULL; kable(Xmtx2, "simple")
```

1	1	0	1	0	1	0	0	0
1	1	0	0	1	0	0	1	0
1	1	0	0	0	0	0	0	0
1	1	0	1	0	1	0	0	0
1	1	0	0	1	0	0	1	0
1	1	0	0	0	0	0	0	0
1	0	1	1	0	0	1	0	0
1	0	1	0	1	0	0	0	1

1	0	1	0	0	0	0	0	0
1	0	1	1	0	0	1	0	0
1	0	1	0	1	0	0	0	1
1	0	1	0	0	0	0	0	0
1	0	0	1	0	0	0	0	0
1	0	0	0	1	0	0	0	0
1	0	0	0	0	0	0	0	0
1	0	0	1	0	0	0	0	0
1	0	0	0	1	0	0	0	0
1	0	0	0	0	0	0	0	0

b. Show that the linear trend for one *group* compared to another (say *GroupA* versus *GroupB*) is estimable by showing that $\boldsymbol{L} = \boldsymbol{LH}$, where the Moore-Penrose inverse is used in calculating $\boldsymbol{H}$. First you need to construct $\boldsymbol{L}$. (As a check, you can repeat using SAS's g-inverse in calculating $\boldsymbol{H}$, but you don't need to turn that in.)

You can use SAS PROC IML or R to construct H; 'ginv' is the function in both that uses the MP inverse. So, for example, you can use 'h=ginv(t(x)*x)*t(x)*x'; in SAS PROC IML. Just use the 'x' from 'ai'. Note that 'L=(0 0 0 0 0 0 0 -1 0 1 1 0 -1 0 0 0)' and you will see that LH comes out to be the same. It is possible that there will be some really small numbers that should be 0, but this is just rounding error (in SAS).

```
##     [1]   [2]  [3]   [4]   [5]  [6]  [7]   [8]  [9]  [10]  [11]  [12]  [13]  [14]  [15]  [16]
L1 <- c(0,    0,   0,    0,    0,   0,   0,   -1,   0,   1,    1,    0,   -1,    0,    0,    0)
XtX1 <- t(Xmtx1) %*% Xmtx1
H1 <- MASS::ginv(XtX1) %*% XtX1
kable(round(L1 %*% H1), "simple")
```

0	0	0	0	0	0	0	-1	0	1	1	0	-1	0	0	0

To this end, we can see that $\boldsymbol{L}$ and $\boldsymbol{LH}$ are identical, which means this contrast is estimable. You can verify other contrasts or matrices forms.

c. How would answers in a change in part **a** if an AR(1) structure for $\boldsymbol{R}$ is included? (You do not need to rewrite entire models, just mention what changes).

You can write this as $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_i)$, where R_i has the AR(1) structure.

'lme()'models in 'R' programming is defined in a different way. in 'R' we do not define the $\boldsymbol{R}$ covariance matrix, we define the correlation matrix $AR(1)$ under equal-variance assumption; if there is a violation of equal variance assumption, we use 'weights' argument to adjust variances. Then a $\boldsymbol{R}$ covariance matrix will be build with both 'correlation' and 'weights' (variances).

d. Say that *Time* is treated as continuous (i.e., not included in the CLASS statement in SAS or factor argument in R). Rewrite either the full-rank or less-than-full-rank model (clearly specify which one) and $\boldsymbol{X}$ matrices in **a**. Say the linear term for *Time* is sufficient.

Here we just give a FR model as an example.

$$
\begin{aligned}
&Y_{ij} = \beta_0 + \beta_1 G_1 + \beta_2 G_2 + \beta_3 time + \beta_4 (G_1 \times time) + \beta_5 (G_2 \times time) + b_i + \epsilon_{ij} \\
&b_i \overset{iid}{\sim} N(0, \sigma_b^2) \\
&\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2) \\
&b_i \perp \epsilon_{ij}
\end{aligned}
$$

```
form3 <- formula(y ~ 1 + group + as.integer(time) + group:as.integer(time))
mod_f3 <- model.frame(form3, # Formula
                      # Data frame
                      data = data_s,
                      # Identifier of data records
                      SubjectId = id)
Xmtx3 <- model.matrix(form3, mod_f3,
                      contrasts.arg = list(group = "contr.SAS"))
colnames(Xmtx3) <- NULL; kable(Xmtx3, "simple")
```

1	1	0	1	1	0
1	1	0	2	2	0
1	1	0	3	3	0

1	1	0	1	1	0
1	1	0	2	2	0
1	1	0	3	3	0
1	0	1	1	0	1
1	0	1	2	0	2
1	0	1	3	0	3
1	0	1	1	0	1
1	0	1	2	0	2
1	0	1	3	0	3
1	0	0	1	0	0
1	0	0	2	0	0
1	0	0	3	0	0
1	0	0	1	0	0
1	0	0	2	0	0
1	0	0	3	0	0

e. Say that the times of observation were at 0, 1 and 6 months rather than equally spaced.

f. Would it be appropriate to treat *Time* as a class variable in this case? Explain.

There is no problem in using equally spaced or unequally spaced times for a class variable, since you are estimating levels separately. The unequal spacing does not impose any constraints metrically.Note: for this question I was considering interpretation of the fixed effects. If you are thinking about implications for the covariance structure, just clearly state that in your argument. For example, if you use the standard AR(1) structure, it would not work well with unequally spaced time points.

ii. Suggest a structure for $\boldsymbol{R}_i$ and write it out.

personally I do not assume the R should have subscribt. This error terms should have been remove the fixed effects and random effects as conditionaled residuals. Hence this pattern and parameters related to $\boldsymbol{R}$ should be shared in the population.

There are a couple of possibilities. Since there are only 3 times, it is not very expensive to use the UN structure, since it only adds 6 covariance parameters.

$$\boldsymbol{R}_i = \begin{pmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{06} \\ \sigma_{01} & \sigma_1^2 & \sigma_{56} \\ \sigma_{06} & \sigma_{56} & \sigma_6^2 \end{pmatrix}$$

So in 'R', the SAS UN structure in fact is a Symmetric correlation matrix (with 3 parameters) and a variance vector (with 3 parameter, more precisely one variance, and two correlation parameters)

Another option would be the spatial power structure. It only adds 2 covariance parameters and handles the unequal spacing. It is also referred to as a continuous AR(1) structure (R).

$$\boldsymbol{R}_i = \sigma_\epsilon^2 \begin{pmatrix} 1 & \phi & \phi^6 \\ \phi & 1 & \phi^5 \\ \phi^6 & \phi^5 & 1 \end{pmatrix}$$