Exam-1

Show clear works for partial credits. Typeset your answer. Feel free to ask questions if there is anything unclear.

Qn1: State the main characteristics that one should take into account while running a longitudinal data analysis.

Qn2: Derive the contrast for equality of the difference between the average response at occasions 2 through n and the baseline value in the two groups when the regression coefficient is $\boldsymbol{\kappa} = (\alpha, \theta, \beta_2, \ldots, \beta_n, \gamma_2, \ldots, \gamma_n)$ for the model $y_{i,j,g} = \alpha + \theta g + \beta_j + \gamma_j * g + \epsilon_{i,j}$ for $j = 1, \ldots, n$ and $g = 0, 1$ with $\beta_1 = \gamma_1 = 0$. (For reference 'the average response at occasions 2 through n and the baseline value' is discussed in 199 of the slides pdf. However, the contrast stated in the slide will change here and should be expressed based on $\boldsymbol{\kappa}$ as in the practice problem for AUCMB.)

Answer 1 *We have the expectation as* $\mu_{j,g} = \alpha + g\theta + \beta_j + g\gamma_j$*.*

We formula for 'the average response at occasions 2 through n *and the baseline value' is* $\frac{1}{n-1}\sum_{j=2}^{n}\mu_{j,g} - \mu_{1,g} = \frac{1}{n-1}\sum_{j=2}^{n}\{\alpha + g\theta + \beta_j + g\gamma_j\} - \mu_{1,g} = \frac{1}{n-1}\beta_j + g\frac{1}{n-1}\gamma_j = E_g$ *(say).*

Then, our inference is based on $E_1 - E_0 = \frac{1}{n-1}\gamma_j$*. Thus, we need the contrast* $\boldsymbol{\ell}$*, such that for regression coefficients* $\kappa = (\alpha, \theta, \beta_2, \ldots, \beta_n, \gamma_2, \ldots, \gamma_n)$*,* $\boldsymbol{\ell}^T\kappa = E_1 - E_0$*. In this case,* $\boldsymbol{\ell}^T\kappa = \frac{1}{n-1}(\gamma_2 + \ldots + \gamma_n)$*. Then* $\boldsymbol{\ell} = (0, 0, 0, \ldots, 0, \frac{1}{n-1}, \ldots, \frac{1}{n-1})$ *is a* $2n \times 1$ *vector where the last* $n-1$ *values are* $\frac{1}{n-1}$*, while the rest are 0.*

Qn3: What are the appropriate covariance models for Orthodontic Measurements on Children data based on the observation times? First, examine the covariance by itself. Later fit the linear mixed model from Qn 2 with all possible covariance structures that are reasonable and compare either using LRT or AIC and get the most suited model. While fitting the linear mixed with random effects, consider adding both a random intercept and slope with respect to time.

Answer 2

```
library(tidyr)


#uploading the data
```

```
data <- read.table("orthodontic.txt", quote="\"", comment.char="")



#converting the data from wide to long
data_long <- gather(data, age, growth, V3:V6, factor_key=TRUE)
data_long <- data_long[order(data_long$V1), ]
data_long$age <- as.character(data_long$age)
data_long$age[grep("V3", data_long$age)] = "8"
data_long$age[grep("V4", data_long$age)] = "10"
data_long$age[grep("V5", data_long$age)] = "12"
data_long$age[grep("V6", data_long$age)] = "14"

colnames(data) <- c("ID", "Gender", "Age 8", "Age 10",
                    "Age 12", "Age 14")

colnames(data_long) <- c("ID", "Gender", "Age", "growth")
##########The covariance matrix:

cov_mat <- data.frame(cov(data[,3:6]))
colnames(cov_mat) <- c("Age 8", "Age 10", "Age 12", "Age 14")
rownames(cov_mat) <- c("Age 8", "Age 10", "Age 12", "Age 14")
knitr::kable(cov_mat, digits = 3, caption = "Covariance Matrix")

##########The correlation matrix:
cor_mat <- data.frame(cor(data[,3:6]))
colnames(cor_mat) <- c("Age 8", "Age 10", "Age 12", "Age 14")
rownames(cor_mat) <- c("Age 8", "Age 10", "Age 12", "Age 14")
knitr::kable(cor_mat, digits = 3, caption = "Correlation Matrix")
```

We can consider the following covariance models: 1) unstructured, 2) linear mixed with time random effect.

Models like autoregressive (AR), moving average (MA), and autoregressive moving average (ARMA) are also appropriate as the data is balanced, and the observation times are gridded. However, none of these may work, since the correlation does not decrease with a larger separation in time.

On the other hand, the Gaussian or exponential kernel-based covariances could have been appropriate too if time is considered a continuous predictor. Technically, one can implement Gaussian or exponential kernel-based covariances even for factor-valued time. But they make less sense statistically as the time is fed *as a continuous input while forming the Gaussian or exponential kernel-based covariances.*

```
###################Time as factor##############################

#Unstructured correlation using generalized least square in nlme
library(nlme)
modnlme <- gls(growth ~ as.factor(Age)*Gender, data = data_long,
           correlation = corSymm(form = ~ 1 | ID))

##########linear mixed model with random intercept#################
mod <- lme(growth ~ as.factor(Age)*Gender, random = ~1|ID, data = data_long)

mod.AR1 <- update(mod, correlation = corAR1()) #AR1 correlation structure
mod.MA1 <- update(mod, corr = corARMA(p = 0, q = 1)) #MA1 correlation strucutre
mod.ARMA11 <- update(mod, corr = corARMA(p = 1, q = 1)) #ARMA(1,1) correlation structure

anova(modnlme, mod.AR1)
anova(modnlme, mod.MA1)
anova(modnlme, mod.ARMA11)



anova(mod.AR1, mod.ARMA11)
anova(mod.MA1, mod.ARMA11)



###Likelihood comparison is not conclusive
AIC(modU, mod.AR1, mod.MA1, mod.ARMA11)

###################Time as continuous predictor (not needed for this Qn)############################
library(nlme)
modnlme <- gls(growth ~ Age*Gender, data = data_long,
               correlation = corSymm(form = ~ 1 | ID))

##########linear mixed model with random intercept#################
mod <- lme(growth ~ Age*Gender, random = ~1|ID, data = data_long)
```

```
##########linear mixed model with random intercept + slope#################
modA <- lme(growth ~ Age*Gender, random = ~1+Age|ID, data = data_long)

modExp <- update(mod,correlation = corExp(form = ~ Age))
modGaus <- update(mod,correlation = corGaus(form = ~ Age))

anova(modnlme, modExp)
anova(modnlme, modGaus)
anova(modnlme, modA)
```

Some important notes on the above code:

- *The base model used in* `update` *function only includes a random intercept as the time-varying dependence is incorporated in AR, MA, or ARMA structure. Although this may not be the most appropriate base model always, it is a reasonable default for all cases.*
- *The* `anova` *based LRT comparison is applied only for the cases of nested models.*

Qn4: Compare linear mixed model and two-stage analysis for the Orthodontic Measurements on Children data to study the effect of 'sex' on 'dental growth'. Considering g in Qn2 as 'sex', compute the contrast-based effect of 'sex' and test for its significance.

<u>Answer</u> 3 *Based on the AIC comparison, mod.AR1 has the smallest AIC.*

```
    ###################Testing the contrast based on AR1 model
## set up contrast (linear comb. of coefficients)
ct <- c(0,0,0,0,0,1/3,1/3,1/3)   #This is the contrast from E_1-E_0 assuming Week 1 as time 1

coef <- fixef(mod.AR1)

m <- sum(coef * ct)  ## mean of contrast
v <- t(ct) %*% vcov(mod.AR1) %*% ct   ## variance of contrast
stder <- sqrt(as.numeric(v))      ## standard error
tstat <- m/stder      ## t statistic

##########For large sample##################
2*pnorm(abs(tstat), lower.tail=FALSE)
```

```
############For small sample#############
#number of parameters for mean
par1 <- length(coef)
#number of parameters for random effect variance
par2 <- sum(upper.tri(getVarCov(mod.AR1,type = c("random.effects")), diag=T))



#number of parameters for mean + number of parameters for variance
number_of_parameters_estimated <- par1 +par2
error_df = nrow(data_long) - number_of_parameters_estimated
2*pt(abs(tstat), df=error_df, lower.tail = FALSE)
```

Hence, the effect of 'sex' is not significant.

Now we consider the two-stage model:

```
res.list <- lmList(growth ~ Age | ID, data=data_long)

# extract the estimated model coefficients (intercepts and slopes) and the corresponding variance-co
b <- lapply(res.list, coef)
V <- lapply(res.list, vcov)

estm <- rep(c("intercept","slope"), length(b))
subj <- rep(names(b), each=2)

#one long vector with the model coefficients and the corresponding block-diagonal variance-covarianc
library(metafor)
b <- unlist(b)
V <- bldiag(V)



#multivariate meta-analysis with the model coefficients
#The V matrix contains the variances and covariances of the sampling errors. We also allow for heter
Sex <- dummy(as.factor(rep(data$Gender, each=2)))
estmd <- dummy(as.factor(estm))

Xmat <- cbind("Intercept"=1-estmd, "Slope"=estmd, "Inte_Male"=(1-estmd)*Sex, "Sl_Male"=estmd * Sex)
```

```
colnames(Xmat) <- c("Intercept", "Slope", "Intercept_male", "Slope_male")
```

```
res2 <- gls(unlist(b) ~ Xmat - 1, correlation = corSymm(form = ~ 1 | subj))
summary(res2)
```

Here 'XmatSlope_male' is of interest and it is not significant. The design matrix is formed following A_i of Page 375 on the slide.

Qn5: Why random effect distribution is assumed to have 0 expectation? Why only time-varying predictors should be considered as covariates in individual-specific random effects?

Answer 4 Random effect expectation

The zero expectation of the random effects is by choice and it is only appropriate when a fixed effect term of the same predictor is also included in the model to estimate the population average effect. Like the following model, the random effect term $b_{2,i}$ should not be assumed to have zero expectation, but the expectation of $b_{1,i}$ can be assumed to be zero as α will capture population average intercept.

$y_{i,j} = \alpha + b_{1,i} + b_{2,i}t_{i,j} + \epsilon_{i,j}$

Time-varying predictors

Take the two models

1: $y_{i,j,g} = \alpha + \theta g + \beta t_{i,j} + b_{1,i} + b_{2,i}t_{i,j} + b_{3,i}g + \epsilon_{i,j,g}$ *with two groups* $g = 0$ *and* $g = 1$.

2: $y_{i,j,g} = \alpha + \theta g + \beta t_{i,j} + b_{1,i} + b_{2,i}t_{i,j} + \epsilon_{i,j,g}$

A given subject can only belong to one of the two groups. $b_{1,i}, b_{2,i}, b_{3,i}$ *are all subject-specific terms. If subject 1 is group '0' its individual-specific intercept is* $b_{1,1}^{(1)}$ *and if subject 2 is group '1', its individual-specific intercept is* $b_{1,2}^{(1)} + b_{3,2}^{(1)}$ *under model 1. Under model 2, these terms are* $b_{1,1}^{(2)}$ *and* $b_{1,2}^{(2)}$ *respectively and thus* $b_{1,1}^{(2)} = b_{1,1}^{(1)}$ *and* $b_{1,2}^{(2)} = b_{1,2}^{(1)} + b_{3,2}^{(1)}$ *without any information loss. Since individual-specific specific treatment effects are usually not of any direct interest, model 2 and model 1 are equivalent.*