Homework 3 Solutions

Alexander McLain

This assingment is slightly different than the one I assigned, however, it has all the main aspects the same. However, the points for each problem are not accurate. See the SAS solution for detailed information on the points and grading for each problem.

1. The dataset `MITgrowth.csv` on the course website data are from a prospective study on body fat accretion in a cohort of 162 girls from the MIT Growth and Development Study. The study was designed to look at changes in percent body fat in girls before and after menarche. All subjects had to be pre-menarche and non-obese to enter the study. Observations were taken annually until 4 years after menarche. At each observation percent body fat was measured.

Two time-scales are included: age, and time since menarche (which can be negative). Time since menarche is the more biologically relevant time scale to use. The variables (in order) are: *Subject ID, Current Age (years), Age at Menarche (years), time relative to Menarche (years), Percent Body Fat.*

First, we'll read in the data. Note that the first row of data does not have the variable names. So, we'll have to use the `header = FALSE` option then fill the names in.

```
library(tidyverse)
MITgrowth <- read.csv("MITgrowth.csv",header = FALSE, na.strings = "",
                      stringsAsFactors = FALSE)
names(MITgrowth) <- c("ID","Age","Age.men","time.r.men","Per.BF")
str(MITgrowth)
```

```
## 'data.frame':    1049 obs. of  5 variables:
##  $ ID        : int  1 1 1 1 1 1 2 2 2 2 ...
##  $ Age       : num  9.32 10.33 11.24 12.19 13.24 ...
##  $ Age.men   : num  13.2 13.2 13.2 13.2 13.2 ...
##  $ time.r.men: num  -3.87 -2.86 -1.95 -1 0.05 1.05 -4.44 -3.2 -2.25 -0.51 ...
##  $ Per.BF    : num  7.24 15.05 13 22.82 10.23 ...
```

```
head(MITgrowth)
```

```
##   ID   Age Age.men time.r.men Per.BF
## 1  1  9.32   13.19      -3.87   7.24
## 2  1 10.33   13.19      -2.86  15.05
## 3  1 11.24   13.19      -1.95  13.00
## 4  1 12.19   13.19      -1.00  22.82
## 5  1 13.24   13.19       0.05  10.23
## 6  1 14.24   13.19       1.05  20.56
```

a. **(10 points) Produce a spaghetti plot of the data using Age and then time relative to menarche.**

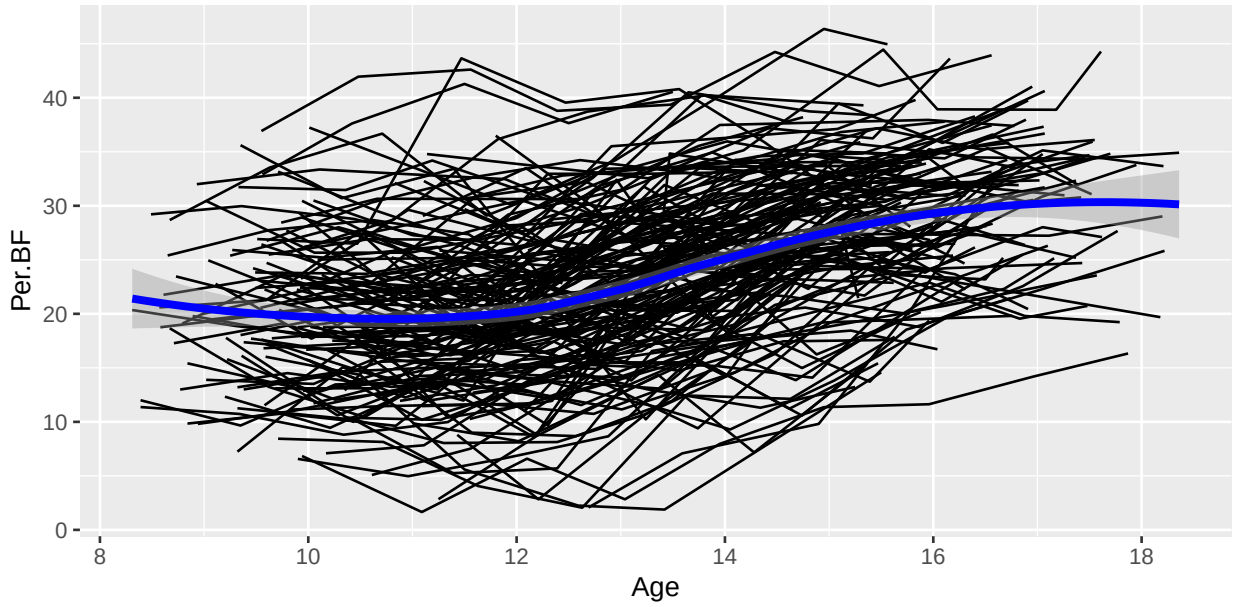
Which time scale is best to answer the study question? Well, there are two "time" " scales in this data: Age and time relative to menarche. Let's look to see which is more closely related to the outcome. Adding a mean line was not necessary, but here I do. Since the data are unbalenced, i'll add the mean line using the "loess" we used for the Six Cities dataset.

First, we'll look at age.

```
p <- ggplot(data = MITgrowth, aes(x = Age, y = Per.BF, group = ID))
p + geom_line() +
  geom_smooth(aes(group = 1), method = "loess", color = "blue", size = 1.5)
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

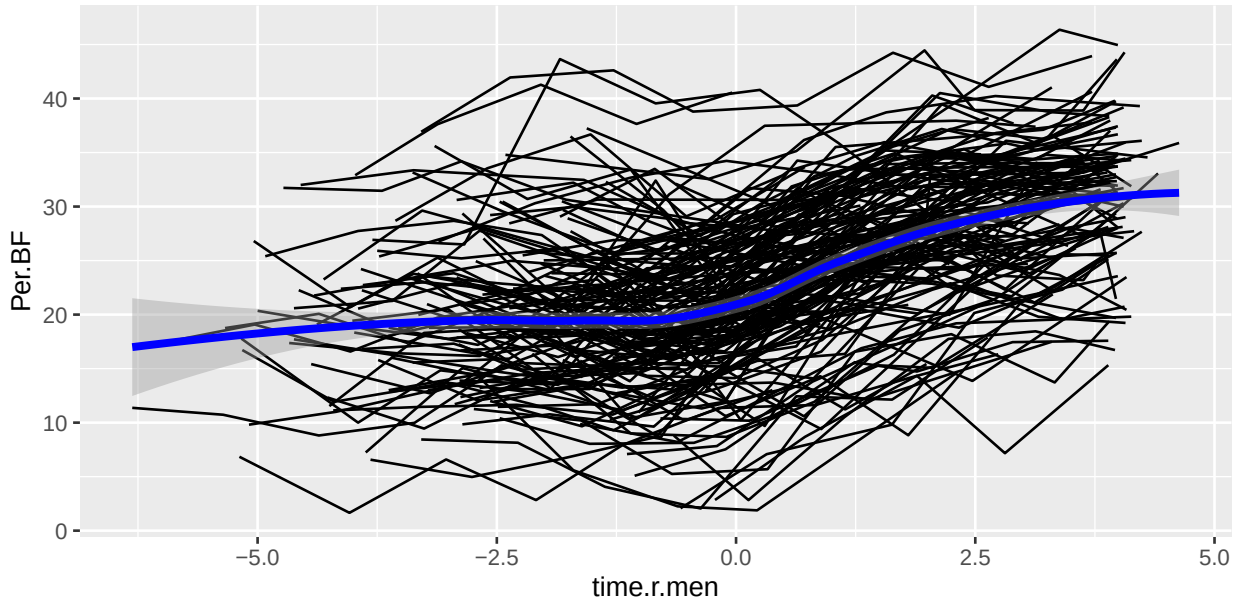
```
## `geom_smooth()` using formula = 'y ~ x'
```



Second, we'll look at time relative to menarche.

```
p <- ggplot(data = MITgrowth, aes(x = time.r.men, y = Per.BF, group = ID))
p + geom_line() +
  geom_smooth(aes(group = 1), method = "loess", color = "blue", size = 1.5)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



b. Which time scale appears to have a stronger relationship with percent body fat? Which time scale is best to answer the study question? Comment on the possible parametric methods that could be used to include time in the model.

There's not a big difference in which time variable appears to be more closely related with the outcome, but I would probably say time relative to menarche looks more closely related. Also, since this study is interested in look at changes in percent body fat in girls before and after menarche it alings with the study objectives and is what I'll use going forward.

The possible parametric methods are a linear spline (at time.r.men = 0). This makes a lot of sense since it aligns with the objectives of the study. We could possibly add a log-linear, or square-root transformations of time before and after menarche.

c. **(10 points) Fit a model with the time effect you found to be most appropriate in (a), with randomly varying intercepts and slopes.**

Here i'm going to use the linear spline model. First, I need to add a variable that will allow for a different linear trend before and after menarche.

```
MITgrowth <- MITgrowth %>% mutate( time.r.men.pos = if_else(time.r.men>0, time.r.men, 0) )
head( MITgrowth )
```

```
##   ID   Age Age.men time.r.men Per.BF time.r.men.pos
## 1  1  9.32   13.19      -3.87   7.24           0.00
## 2  1 10.33   13.19      -2.86  15.05           0.00
## 3  1 11.24   13.19      -1.95  13.00           0.00
## 4  1 12.19   13.19      -1.00  22.82           0.00
## 5  1 13.24   13.19       0.05  10.23           0.05
## 6  1 14.24   13.19       1.05  20.56           1.05
```

Now, I'm going to fit a model with random varying intercepts and slopes.

```
library(lme4)
library(lmerTest)
```

```
LMM_formula <- Per.BF ~ time.r.men + time.r.men.pos + (1 + time.r.men|ID)
LMM_linear_sp_int_slp <- lmer( formula = LMM_formula , data = MITgrowth)
summary(LMM_linear_sp_int_slp)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: LMM_formula
##    Data: MITgrowth
##
## REML criterion at convergence: 6085.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.8202 -0.6036 -0.0110  0.6060  3.3196
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  ID       (Intercept) 37.9147  6.1575
##           time.r.men   0.5333  0.7303   -0.24
##  Residual             10.4313  3.2298
## Number of obs: 1049, groups:  ID, 162
##
## Fixed effects:
##                Estimate Std. Error       df t value Pr(>|t|)
## (Intercept)     20.9902     0.5205 193.2797  40.324  < 2e-16 ***
## time.r.men       0.5095     0.1291 461.1871   3.946  9.2e-05 ***
## time.r.men.pos   2.3535     0.1867 868.3733  12.605  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) tm.r.m
## time.r.men   0.130
## tim.r.mn.ps -0.301 -0.804
```

i. **(5 points) What is the estimated variance of the random intercepts?**

The estimated variance of the random intercepts is 37.916.

ii. **(5 points) What is the estimated variance of the random slopes?**

The estimated variance of the random slopes is 0.533.

iii. **(5 points) What is the estimated correlation between the random intercepts and slopes?**

The estimated correlation between the random intercepts and slopes is -0.24.

d. **(10 points) Print the v and vcorr matrix from the model in (c). Comment on the heterogeneity and pattern in the correlations over time.**

To get the `v` and `vcorr` matrices we need to fit the model with `lme` instead of `lmer`.

```
library(nlme)
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:lme4':
##
##     lmList
```

```
## The following object is masked from 'package:dplyr':
```

```
##
##     collapse
```

```
library(Matrix)

LMM_formula_fixed <- Per.BF ~ time.r.men + time.r.men.pos
LMM_formula_random <- ~ time.r.men|ID

LMM_linear_sp_int_slp_lme <- lme(LMM_formula_fixed, random = LMM_formula_random, data = MITgrowth)
summary(LMM_linear_sp_int_slp_lme)
```

```
## Linear mixed-effects model fit by REML
##   Data: MITgrowth
##        AIC      BIC    logLik
##   6099.246 6133.915 -3042.623
##
## Random effects:
##  Formula: ~time.r.men | ID
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev   Corr
## (Intercept) 6.157427 (Intr)
## time.r.men  0.730290 -0.244
## Residual    3.229749
##
## Fixed effects:  list(LMM_formula_fixed)
##                    Value Std.Error  DF  t-value p-value
## (Intercept)    20.990226 0.5205340 885 40.32441   0e+00
## time.r.men      0.509538 0.1291432 885  3.94553   1e-04
## time.r.men.pos  2.353453 0.1867092 885 12.60491   0e+00
##  Correlation:
##                (Intr) tm.r.m
## time.r.men      0.130
## time.r.men.pos -0.301 -0.804
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -2.8202061 -0.6036131 -0.0110199  0.6060221  3.3196620
##
## Number of Observations: 1049
## Number of Groups: 162
```

We can see that the results are the same as when `lmer` is used. (`lme` and `lmer` were developed by the same authors.)

Now, let's get the `v` and `vcorr` matrices.

```
(V <- getVarCov(LMM_linear_sp_int_slp_lme,type="marginal",individual=1) )
```

```
## ID 1
## Marginal variance covariance matrix
##        1      2      3      4      5      6
## 1 64.821 51.197 48.321 45.318 42.000 38.839
## 2 51.197 58.980 46.163 43.672 40.919 38.297
## 3 48.321 46.163 54.650 42.189 39.946 37.809
## 4 45.318 43.672 42.189 51.072 38.929 37.299
## 5 42.000 40.919 39.946 38.929 48.237 36.736
```

```
## 6 38.839 38.297 37.809 37.299 36.736 46.630
##   Standard Deviations: 8.0511 7.6799 7.3926 7.1465 6.9453 6.8286
```

```
(Vcorr <- cov2cor(simplify2array(getVarCov(LMM_linear_sp_int_slp_lme,type="marginal",individual=1)[[1]]))
```

```
##           1         2         3         4         5         6
## 1 1.0000000 0.8280105 0.8118661 0.7876406 0.7511065 0.7064482
## 2 0.8280105 1.0000000 0.8131037 0.7957219 0.7671565 0.7302645
## 3 0.8118661 0.8131037 1.0000000 0.7985703 0.7780067 0.7489710
## 4 0.7876406 0.7957219 0.7985703 1.0000000 0.7843208 0.7643176
## 5 0.7511065 0.7671565 0.7780067 0.7843208 1.0000000 0.7745758
## 6 0.7064482 0.7302645 0.7489710 0.7643176 0.7745758 1.0000000
```

Main factors: (i) the variance appears to be slightly decreasing over time, which is unusual in public health longitudinal data, (ii) the correlations decrease slightly with increasing time separation, and (iii) the correlations that are 1 unit apart decrease as the subjects get older.

e. **(10 points) Fit a model with only randomly varying intercepts. What do you think about this model versus the previous model? Should this be used? Why?**

First, let's fit a model with only randomly varying intercepts and compare the fit to the intercepts and slopes model.

```
LMM_formula <- Per.BF ~ time.r.men + time.r.men.pos + (1|ID)
LMM_linear_sp_int <- lmer( formula = LMM_formula , data = MITgrowth)
anova(LMM_linear_sp_int,LMM_linear_sp_int_slp)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: MITgrowth
## Models:
## LMM_linear_sp_int: LMM_formula
## LMM_linear_sp_int_slp: LMM_formula
##                       npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## LMM_linear_sp_int        5 6160.2 6185.0 -3075.1   6150.2
## LMM_linear_sp_int_slp    7 6094.8 6129.5 -3040.4   6080.8 69.428  2  8.394e-16
##
## LMM_linear_sp_int
## LMM_linear_sp_int_slp ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The likelihood ratio test is highly significant, which indicates that the more complex model (the one with random slopes) fits significantly better than the smaller model (the one with no random slopes). AIC and BIC tell the same story. This indicates that the random slope should be left in the model.

f. **(20 points) Give a full description of your findings. Include interpretations of at least two regression coefficients in your description. All descriptions should be in context of the study and the goals of the study.**

First, I'm going to get confidence intervals for all of the β coefficients, and then print them with the actual values.

```
CI_lims <- confint(LMM_linear_sp_int_slp , parm = "beta_")
```

```
## Computing profile confidence intervals ...
```

```
round( cbind( fixef(LMM_linear_sp_int_slp), CI_lims) , 3)
```

```
##                       2.5 % 97.5 %
```

```
## (Intercept)    20.990 19.968 22.012
## time.r.men      0.510  0.256  0.763
## time.r.men.pos  2.353  1.988  2.720
```

The study finds that there is a significant increase in body fat percentage before Menarche. Specifically, before Menarche we expect an increase of 0.409 percentage points for each year they get older. After Menarche, there is a significantly higher increase in body fat percentage than before Menarche. Specifically, after Menarche girls body fact increases 2.055 percentage points more per year than before Menarche. After Menarche girls body fat increases 2.46 $(0.409 + 2.055)$ percentage points per year.