ST437/537 – HW #07

Arnab Maity

Due date: April 16, 2019

Instructions

Please follow the instructions below when you prepare and submit your assignment.

- **Include a cover-page** with your homework. It should contain

i. Full name,
ii. Course#: ST 437/537 and
iii. HW-#
iv. Submission date

- Assignments should be submitted in class on the date specified ("due date").
- Neatly typed or hand-written solution on standard letter-size papers (stapled on the top-left corner) should be submitted. **All R code/output should be well commented, with relevant outputs highlighted.**
- **Always staple (upper left corner) your homework <u>before coming to class.</u> Ten percent points will be deducted otherwise.**
- When you solve a particular problem, do not only give the final answer. Instead **show all your work** and the steps you used (with proper explanation) to arrive at your answer to get full credit.
- **DO NOT** give printouts of whole dataset or matrices. Present only the relevant output when answering a question.

Problems

Solve the following problems. You may use `R` for these problems unless I specifically instruct otherwise.

DO NOT give printouts of whole dataset or matrices. Present only the relevant output/graphs when answering a question.

Problem 1

Refer to the Six Cities Air Pollution data (Applied Longitudinal Data Analysis by Fitzmaurice, Laird and Ware http://www.hsph.harvard.edu/fitzmaur/ala/ (http://www.hsph.harvard.edu/fitzmaur/ala/)). See also Example 2 in the [Introduction] (https://www.stat.ncsu.edu/people/maity/courses/st537-S2019/Lecture07_LDA_Introduction.html) lecture for data description.

The dataset is in the file [airpollution_all.txt] (../data/airpollution_all.txt)

```
library(lattice)
library(latticeExtra)
```

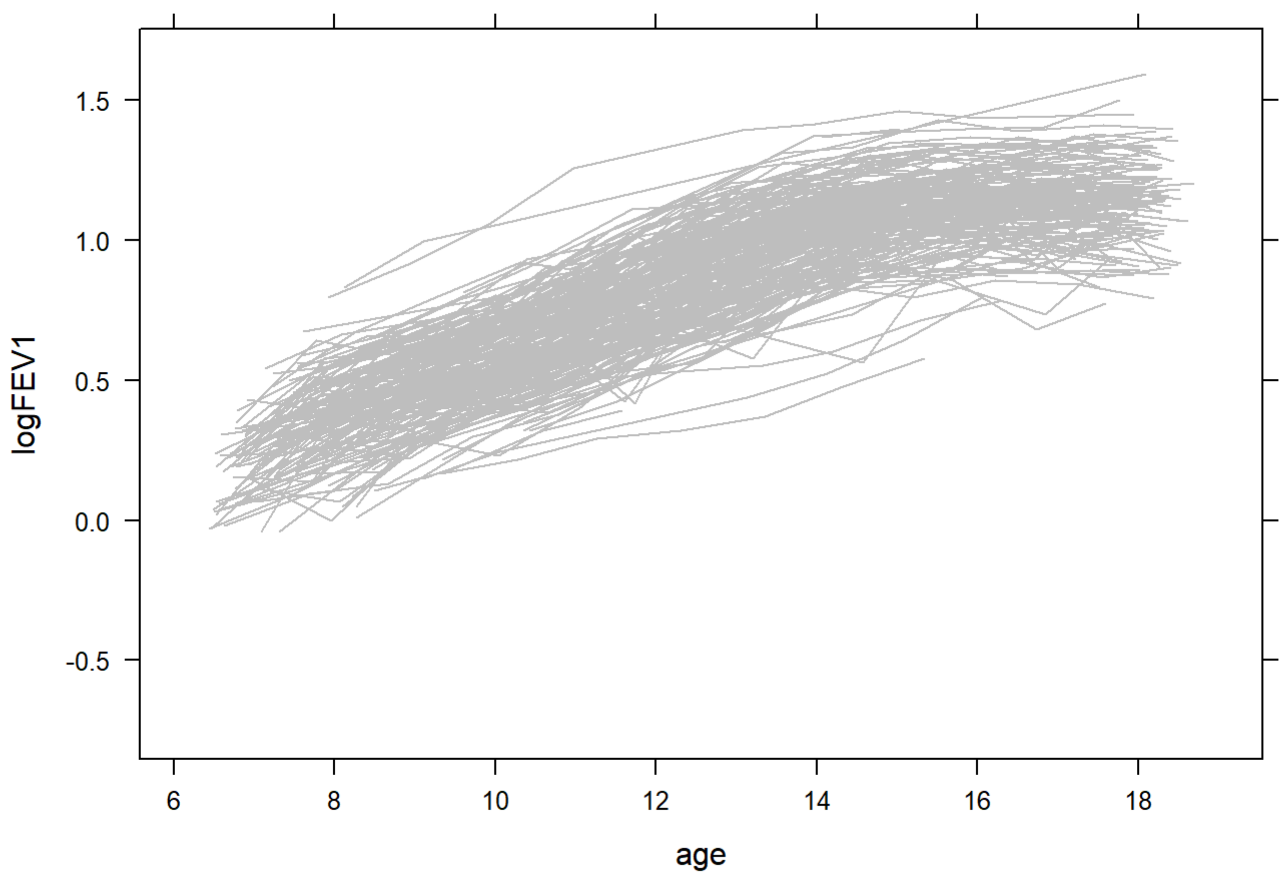
```
## Loading required package: RColorBrewer
```

```
# read data
pollution <- read.table("../data/airpollution_all.txt")

colnames(pollution)=c("id", "Height", "age", "height_base", "age_base", "logFEV1")
head(pollution)
```

```
##   id Height     age height_base age_base logFEV1
## 1  1   1.20  9.3415         1.2   9.3415 0.21511
## 2  1   1.28 10.3929         1.2   9.3415 0.37156
## 3  1   1.33 11.4524         1.2   9.3415 0.48858
## 4  1   1.42 12.4600         1.2   9.3415 0.75142
## 5  1   1.48 13.4182         1.2   9.3415 0.83291
## 6  1   1.50 15.4743         1.2   9.3415 0.89200
```

```
xyplot(logFEV1 ~ age, data = pollution, groups = id, type="l", col="grey")
```



(a) Start with fitting a random intercept model:

$$logFEV1_{ij} = \beta_{0i} + \beta_1 age_{ij} + \beta_2 Height_{ij} + e_{ij},$$

where $\boldsymbol{e}_i \sim N(0, \boldsymbol{I})$, $\beta_{0i} = \beta_0 + b_{0i}$ and $b_{0i} \sim N(0, D_{11})$.

Find the estimates of all the model parameters (regression coefficients and variance components).

```
library(nlme)
fit1 = lme(logFEV1 ~ age + Height, data=pollution, random = ~ 1|id)
summary(fit1)
```

```
## Linear mixed-effects model fit by REML
##  Data: pollution
##         AIC       BIC  logLik
##   -4472.32 -4444.338 2241.16
##
## Random effects:
##  Formula: ~1 | id
##         (Intercept)   Residual
## StdDev:   0.1056158 0.06368607
##
## Fixed effects: logFEV1 ~ age + Height
##                  Value  Std.Error   DF   t-value p-value
## (Intercept) -1.8584598 0.03072320 1692 -60.49044       0
## age          0.0197742 0.00131214 1692  15.07017       0
## Height       1.6186518 0.03013247 1692  53.71786       0
##  Correlation:
##        (Intr) age
## age     0.834
## Height -0.961 -0.935
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -5.87179979 -0.51990855  0.07062222  0.59867581  2.82890024
##
## Number of Observations: 1994
## Number of Groups: 300
```

```
## Regression coefs
fit1$coefficients$`fixed`
```

```
## (Intercept)         age      Height
## -1.85845979  0.01977417  1.61865179
```

```
## Variance components: D
getVarCov(fit1)
```

```
## Random effects variance covariance matrix
##             (Intercept)
## (Intercept)    0.011155
##   Standard Deviations: 0.10562
```

```
## Variance components: sigma^2
sigma(fit1)^2
```

```
## [1] 0.004055915
```

(b) Write a model (as we have done in part (a)) where we include random coefficients for both intercept and slope of height, but not for age assuming that the random effects are independent.

Fit this model and find the estimates of all the model parameters (regression coefficients and variance components).

Model:

$$logFEV1_{ij} = \beta_{0i} + \beta_1 age_{ij} + \beta_{2i} Height_{ij} + e_{ij},$$

where $\boldsymbol{e}_i \sim N(0, \boldsymbol{I})$, $\beta_{0i} = \beta_0 + b_{0i}$ and $\beta_{2i} = \beta_2 + b_{2i}$. Here we have

$$\begin{pmatrix} b_{0i} \\ b_{2i} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} D_{11} & 0 \\ 0 & D_{22} \end{pmatrix}\right].$$

Combining, we have the model

$$logFEV1_{ij} = \beta_0 + \beta_1 age_{ij} + \beta_2 Height_{ij} + b_{0i} + b_{2i} Height_{ij} + e_{ij}.$$

```
fit2 = lme(logFEV1 ~ age + Height, data=pollution,
  random = list(id = pdDiag(~Height)))
summary(fit2)
```

```
## Linear mixed-effects model fit by REML
##  Data: pollution
##         AIC       BIC   logLik
##   -4493.282 -4459.703 2252.641
##
## Random effects:
##  Formula: ~Height | id
##  Structure: Diagonal
##         (Intercept)     Height   Residual
## StdDev:  0.07549434 0.05387405 0.06271315
##
## Fixed effects: logFEV1 ~ age + Height
##                  Value   Std.Error   DF   t-value p-value
## (Intercept) -1.8709869 0.030328201 1692 -61.69132       0
## age          0.0193256 0.001301116 1692  14.85310       0
## Height       1.6308339 0.030167597 1692  54.05913       0
##  Correlation:
##        (Intr) age
## age     0.839
## Height -0.966 -0.927
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -5.95366206 -0.52259826  0.06687414  0.59033619  2.60902831
##
## Number of Observations: 1994
## Number of Groups: 300
```

```
## Regression coefs
fit2$coefficients$`fixed`
```

```
## (Intercept)         age      Height
## -1.87098686  0.01932561  1.63083391
```

```
## Variance components: D
getVarCov(fit2)
```

```
## Random effects variance covariance matrix
##             (Intercept)    Height
## (Intercept)   0.0056994 0.0000000
## Height        0.0000000 0.0029024
##   Standard Deviations: 0.075494 0.053874
```

```
## Variance components: sigma^2
sigma(fit2)^2
```

```
## [1] 0.003932939
```

(c) Write a model (as we have done in part (a)) where we include random coefficients for both intercept and slope of height, but not for age, assuming that the random effects are dependent with an unstructured covariance matrix.

Fit this model and find the estimates of all the model parameters (regression coefficients and variance components).

Model:

$$logFEV1_{ij} = \beta_{0i} + \beta_1 age_{ij} + \beta_{2i} Height_{ij} + e_{ij},$$

where $\boldsymbol{e}_i \sim N(0, \boldsymbol{I})$, $\beta_{0i} = \beta_0 + b_{0i}$ and $\beta_{2i} = \beta_2 + b_{2i}$. Here we have

$$\begin{pmatrix} b_{0i} \\ b_{2i} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} D_{11} & D_{12} \\ D_{12} & D_{22} \end{pmatrix}\right].$$

Combining, we have the model

$$logFEV1_{ij} = \beta_0 + \beta_1 age_{ij} + \beta_2 Height_{ij} + b_{0i} + b_{2i} Height_{ij} + e_{ij}.$$

```
fit3 = lme(logFEV1 ~ age + Height, data=pollution, random = ~ Height|id)
summary(fit3)
```

```
## Linear mixed-effects model fit by REML
##  Data: pollution
##         AIC       BIC   logLik
##   -4577.172 -4537.997 2295.586
##
## Random effects:
##  Formula: ~Height | id
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev     Corr
## (Intercept) 0.29190910 (Intr)
## Height      0.19588375 -0.936
## Residual    0.05819438
##
## Fixed effects: logFEV1 ~ age + Height
##                  Value  Std.Error   DF   t-value p-value
## (Intercept) -1.9033332 0.03499437 1692 -54.38969       0
## age          0.0187597 0.00124855 1692  15.02515       0
## Height       1.6575527 0.03188128 1692  51.99142       0
##  Correlation:
##        (Intr) age
## age     0.709
## Height -0.962 -0.848
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -6.49207504 -0.49664489  0.08001312  0.56603427  2.90447701
##
## Number of Observations: 1994
## Number of Groups: 300
```

```
## Regression coefs
fit3$coefficients$`fixed`
```

```
## (Intercept)         age      Height
##  -1.9033332   0.0187597   1.6575527
```

```
## Variance components: D
getVarCov(fit3)
```

```
## Random effects variance covariance matrix
##             (Intercept)    Height
## (Intercept)    0.085211 -0.053533
## Height        -0.053533  0.038370
##   Standard Deviations: 0.29191 0.19588
```

```
## Variance components: sigma^2
sigma(fit3)^2
```

```
## [1] 0.003386586
```

(d) Based on the output in the parts above, discuss the following:

i. how estimates of the regression parameters change (or not change),

ii. difference between the covariance structures of the random effects,

iii. create an table with AIC/BIC, and which of the three models you prefer.

The regression coefficients do not change drastically between the three fits. However, the correlation between the random effects in fit3 is -0.93, and thus assumeption of indeoendent random effects (i.e., fit 2) might not be valid here.

The table of AIC/BIC values are below.

```
aic <- AIC(fit1, fit2, fit3)
bic <- BIC(fit1, fit2, fit3)
cbind(aic, bic$BIC)
```

```
##      df       AIC   bic$BIC
## fit1  5 -4472.320 -4444.338
## fit2  6 -4493.282 -4459.703
## fit3  7 -4577.172 -4537.997
```

From the AIC/BIC table, it seems that `fit3` (i.e., dependent random effects) is the best choice here.

(e) Based on the model in your answer in (d)(iii), answer the following questions.

i. Let $Y_{ij} = logFEV_{ij}$. Write the foumula of $E(Y_{ij})$ assuming age_{ij} and $Height_{ij}$ are fixed.

ii. Find $var(Y_{ij})$ and $cov(Y_{ij}, Y_{ik})$.

iii. Estimate the mean of logFEV1, $E(Y_{ij})$, for girls at $age = 12$ and $height = 1.4$ (at that age)

Based on the choice in d(iii), we have the model,

$$logFEV1_{ij} = \beta_0 + \beta_1 age_{ij} + \beta_2 Height_{ij} + b_{0i} + b_{2i} Height_{ij} + e_{ij},$$

where

$$\begin{pmatrix} b_{0i} \\ b_{2i} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} D_{11} & D_{12} \\ D_{12} & D_{22} \end{pmatrix} \right].$$

So, $E(Y_{ij}) = E(\beta_0 + \beta_1 age_{ij} + \beta_2 Height_{ij} + b_{0i} + b_{2i} Height_{ij} + e_{ij}) = \beta_0 + \beta_1 age_{ij} + \beta_2 Height_{ij}$, since $E(b_{0i}) = E(b_{1i} = E(e_{ij}) = 0.$

For part (ii):

$$var(Y_{ij}) = var(\beta_0 + \beta_1 age_{ij} + \beta_2 Height_{ij} + b_{0i} + b_{2i} Height_{ij} + e_{ij})$$

$$= var(b_{0i}) + var(b_{2i}Height_{ij}) + var(e_{ij}) + 2cov(b_{0i}, b_{2i}Height_{ij})$$

$$= D_{11} + D_{22}Height_{ij}^2 + \sigma^2 + 2D_{12}Height_{ij}$$

Similarly

$$cov(Y_{ij}, Y_{ik}) = cov(b_{0i} + b_{2i}Height_{ij}, b_{0i} + b_{2i}Height_{ik})$$

$$= D_{11} + D_{22}Height_{ij}Height_{ik} + 2D_{12}Height_{ij} + D_{12}Height_{ik}$$

For part (iii), we have $age = 12$ and $height = 1.4$, and thus

$$E(Y_{ij}) = (-1.9033332) + (0.0187597)(12) + (1.6575527)(1.4) = 0.642357.$$

Problem 2

Using the same data set used in problem 1, consider the following output and answer the questions that follow.

```
## Linear mixed-effects model fit by REML
##  Data: pollution
##         AIC       BIC   logLik
##   -4577.172 -4537.997 2295.586
##
## Random effects:
##  Formula: ~Height | id
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev     Corr
## (Intercept) 0.29190910 (Intr)
## Height      0.19588375 -0.936
## Residual    0.05819438
##
## Fixed effects: logFEV1 ~ age + Height
##                  Value  Std.Error   DF   t-value p-value
## (Intercept) -1.9033332 0.03499437 1692 -54.38969       0
## age          0.0187597 0.00124855 1692  15.02515       0
## Height       1.6575527 0.03188128 1692  51.99142       0
##  Correlation:
##        (Intr) age
## age     0.709
## Height -0.962 -0.848
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -6.49207504 -0.49664489  0.08001312  0.56603427  2.90447701
##
## Number of Observations: 1994
## Number of Groups: 300
```

(a) Write the mathematical model that is being fit above. Clearly specify all the regression parameters and the variance components.

we have the model,

$$logFEV1_{ij} = \beta_0 + \beta_1 age_{ij} + \beta_2 Height_{ij} + b_{0i} + b_{2i} Height_{ij} + e_{ij},$$

where

$$\begin{pmatrix} b_{0i} \\ b_{2i} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} D_{11} & D_{12} \\ D_{12} & D_{22} \end{pmatrix}\right].$$

(b) Find random effects correlation matrix and variance-covariance matrix (D)

From the “Random effects:” part of the output (see the StdDev column and Corr column), we have

$$D_{11} = (0.29190910)^2 = 0.0852109$$

$$D_{22} = (0.19588375)^2 = 0.0383704$$

$$D_{12} = corr(b_{0i}, b_{2i}) * \sqrt{D_{11}D_{22}} = (-0.936)(0.29190910)(0.19588375) = -0.0535207$$

Thus the covariance matrix is

$$\begin{pmatrix} 0.0852109 & -0.0535207 \\ -0.0535207 & 0.0383704 \end{pmatrix}$$

The correlation matrix is

$$\begin{pmatrix} 1 & -0.936 \\ -0.936 & 1 \end{pmatrix}$$

(c) Estimate the error variance σ^2.

Estimate of σ^2 is $(0.05819438)^2 = 0.0033866$. This is taken from the random effects part of the output (see the StdDev column for Residual)

(d) Let $Y_{ij} = logFEV_{ij}$. Write the foumula of $E(Y_{ij})$ assuming age_{ij} and $Height_{ij}$ are fixed.

Here

$$E(logFEV1_{ij}) = \beta_0 + \beta_1 age_{ij} + \beta_2 Height_{ij}$$

(e) Find $var(Y_{ij})$ and $cov(Y_{ij}, Y_{ik})$.

We have

$$\begin{aligned} var(Y_{ij}) &= var(\beta_0 + \beta_1 age_{ij} + \beta_2 Height_{ij} + b_{0i} + b_{2i} Height_{ij} + e_{ij}) \\ &= var(b_{0i}) + var(b_{2i} Height_{ij}) + var(e_{ij}) + 2cov(b_{0i}, b_{2i} Height_{ij}) \\ &= D_{11} + D_{22} Height_{ij}^2 + \sigma^2 + 2D_{12} Height_{ij} \end{aligned}$$

Similarly

$$\begin{aligned} cov(Y_{ij}, Y_{ik}) &= cov(b_{0i} + b_{2i} Height_{ij}, b_{0i} + b_{2i} Height_{ik}) \\ &= D_{11} + D_{22} Height_{ij} Height_{ik} + D_{12} Height_{ij} + D_{12} Height_{ik} \end{aligned}$$

(f) Consider the following data for one girl:

```
##      id Height    age height_base age_base logFEV1
## 168 24   1.18 6.5216        1.18   6.5216 0.19062
## 169 24   1.23 7.4743        1.18   6.5216 0.42527
```

Find the variance covariance matrix ($V_{2\times 2}$) of logFEV1 for this individual.

Here we have two observations with $Height_{i1} = 1.18$ and $Height_{i2} = 1.23$. From part (e)

$$var(Y_{i1}) = D_{11} + D_{22}Height_{i1}^2 + \sigma^2 + 2D_{12}Height_{i1} = 0.0156865$$

$$var(Y_{i2}) = D_{11} + D_{22}Height_{i2}^2 + \sigma^2 + 2D_{12}Height_{i2} = 0.0149569$$

$$cov(Y_{i1}, Y_{i2}) = D_{11} + D_{22}Height_{i1}Height_{i2} + D_{12}Height_{i1} + D_{12}Height_{i2} = 0.0118872.$$

Thus the covariance matrix is

```
##            1          2
## 1 0.01568654 0.01188716
## 2 0.01188716 0.01495687
```

(g) Give a decomposition of V above into between-subject covariance and within subject covariance matrices. Which part do you think contribute most to V?

Note the we are fitting a model with independent errors, that is

$$cov(e_i) = \sigma^2 I$$

From part(c), this is

```
##             1           2
## 1 0.003386586 0.000000000
## 2 0.000000000 0.003386586
```

This is the within sibjects variance. Thus the between-subject variance is (Total variance) - (within subject variance)

```
##            1          2
## 1 0.01229995 0.01188716
## 2 0.01188716 0.01157029
```

So we can write the decomposition

$$\underbrace{\begin{bmatrix} 0.0156865 & 0.0118872 \\ 0.0118872 & 0.0149569 \end{bmatrix}}_{\text{Total variance}} = \underbrace{\begin{bmatrix} 0.0123 & 0.0118872 \\ 0.0118872 & 0.0115703 \end{bmatrix}}_{\text{Between-subject variance}} + \underbrace{\begin{bmatrix} 0.0033866 & 0 \\ 0 & 0.0033866 \end{bmatrix}}_{\text{Within-subject variance}}$$