BIS 628 HW 3

Joanna Chen

2/15/2020

Question 1 Definition of equivalent linear mixed effects (LME) models: two LME models are said to be equivalent when the mean and the variance- covariance for the response variables are the same between the two models:

(a) Re-write a LME model in which the random intercept and random slope of one covariate (i.e. only two fixed effects and only two random effects are present) have independent correlation structure (G matrix has '0' on the off- diagonal) and the error terms also have independent correlation structure (R matrix has '0s' on the off-diagonal) into an equivalent LME model in which the only random effect present is the slope (there is no random intercept), and there is also an error term.

original:

$$Y_{ij} = \beta_1 + \beta_2 X_{ij} + b_{1i} + b_{2i} X_{ij} + \epsilon_{ij}$$

rewrite:

$$Y_{ij} = \beta_1 + \beta_2 X_{ij} + b_{2i} X_{ij} + \epsilon_{ij}$$

Want to show they are equivalenet. It's equivalent to show that their mean, variance and covariance are the same.
For the original model:
Marginal mean:

$$E(Y_{ij}) = \beta_1 + \beta_2 X_{ij}$$

Variance:

$$\mathrm{Var}(Y_{ij}) = \mathrm{Var}\left(\mathrm{b_{1i}} + \mathrm{b_{2i}X_{ij}} + \epsilon_{\mathrm{ij}}\right) = \mathrm{Var}\left(\mathrm{b_{1i}}\right) + 2X_{ij}\,\mathrm{Cov}\left(b_{1i}b_{2i}\right) + X_{ij}^2\,\mathrm{Var}\left(b_{2i}\right) + \mathrm{Var}\left(\epsilon_{ij}\right) = \sigma_{b1}^2 + 2X_{ij}\,\mathrm{Cov}\left(b_{1,}b_{2i}\right) + X_{ij}^2\sigma_{b2}^2 + \sigma^2$$

Covariance:

$$\mathrm{Cov}\left(Y_{ij}, Y_{ik}\right) = \mathrm{Cov}\left(\mathrm{b_{1i}} + \mathrm{b_{2i}X_{ij}} + \varepsilon_{\mathrm{ij}}, \mathrm{b_{1i}} + \mathrm{b_{2i}X_{ik}} + \varepsilon_{\mathrm{ik}}\right) = \sigma_{b1}^2 + \left(X_{ij} + X_{ik}\right)\mathrm{Cov}\left(b_{1i,}b_{2i}\right) + X_{ij}X_{ik}\sigma_{b2}^2$$

For the rewrited model: Marginal mean remains the same.

$$E(Y_{ij}) = \beta_1 + \beta_2 X_{ij}$$

Variance:

$$\mathrm{Var}(Y_{ij}) = \mathrm{Var}\left(\mathrm{b_{2i}X_{ij}} + \epsilon_{\mathrm{ij}}\right) = X_{ij}^2\,\mathrm{Var}\left(b_{2i}\right) + \mathrm{Var}\left(\epsilon_{ij}\right) = X_{ij}^2\sigma_{b2}^2 + \sigma^2$$

Covariance:

$$\mathrm{Cov}\left(Y_{ij}, Y_{ik}\right) = \mathrm{Cov}\left(\mathrm{b_{2i}X_{ij}} + \varepsilon_{\mathrm{ij}}, \mathrm{b_{2i}X_{ij}} + \varepsilon_{\mathrm{ik}}\right) = X_{ij}X_{ik}\sigma_{b2}^2$$

Two models will be the equivalent model if

$$\sigma_{b1}^2 + 2X_{ij}\,\mathrm{Cov}\left(b_{1,}b_{2i}\right) = 0 \text{ and } \sigma_{b1}^2 + \left(X_{ij} + X_{ik}\right)\mathrm{Cov}\left(b_{1i,}b_{2i}\right) = 0.$$

Since our variance is always positive. The above equations is equivalent to

$$\sigma_{b1}^2 = 0, X_{ij}\,\mathrm{Cov}\left(b_{1i,}b_{2i}\right) = 0 \text{ and } X_{ik}\,\mathrm{Cov}\left(b_{1i,}b_{2i}\right) = 0.$$

(b) For a LME model including fixed effects, random intercept with variance (sigma_b1)^2 and AR(1) error term with variance sigma^2 and consecutive correlation coefficient rho, is there an equivalent linear model without the random intercept but with a new AR(1) error term? Why or Why not?

Model:

$$Y_{ij} = \beta_1 + \beta_2 X_{ij} + b_{1i} + \epsilon_{ij}$$

where

$$\mathrm{b_{1i}} \sim \mathrm{N}\left(0, \sigma^2_{b1}\right)$$

and

$$\epsilon \sim \mathrm{N}(0, \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{pmatrix})$$

and therefore $\epsilon_{ij} \sim \mathrm{N}(0,$ ij-entry of the variance-covariance matrix, denoted by $\sigma_{ij})$.

We have $Var(Y_{ij}) = \sigma^2_{b1} + \sigma^2$ and $Cov(Y_{ij}, Y_{ik}) = \sigma^2_{b1} + \sigma^2_{jk}$ since the error term is no more independent. Therefore,

$$\mathrm{Cov}\left(\mathbf{Y}_i\right) = \begin{pmatrix} \sigma^2_{b1} + \sigma^2 & \sigma^2_{b1} + \sigma^2_{12} & \cdots & \sigma^2_{b1} + \sigma^2_{1n} \\ \sigma^2_{b1} + \sigma^2_{21} & \sigma^2_{b1} + \sigma^2 & \dots & \sigma^2_{b1} + \sigma^2_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2_{b1} + \sigma^2_{n1} & \sigma^2_{b1} + \sigma^2_{n2} & \cdots & \sigma^2_{b1} + \sigma^2 \end{pmatrix}$$

Since $\sigma^2_{b1} \neq 0$, we cannot find a ρ to be able to rewrite the above covariance in the AR(1) form. Therefore, there is not an equivalent linear model without the random intercept but with a new AR(1) error term.

Background for the data set for Question 2 - Question 5

Cystic Fibrosis (CF) is a genetic disease that leads to pulmonary complications and ultimately death. These data represent a subsample of measurements available in a CF Registry database. The dataset has observations on 200 children, with the primary outcome of FEV1 (health status based on how the lung functions). We are interested in examining whether the mean FEV1 change over the follow up is dependent on: (i) child's age (both, cross-sectional, i.e. age at first follow up, and longitudinal age), (ii) child's sex, and (iii) child's F508 genotype. In order to understand the difference between the cross-sectional and longitudinal components of child's age, please, read Chapter 8, Section 8 (p.213-220) and Chapter 9, Sections 5-6 (p.252-258) of "Applied Longitudinal Analysis", Fitzmaurice, Laird, Ware.

Dataset: NewCFkids-SAS.data or NewCFkids-SAS.csv
The dataset is already in a long format, i.e. with repeated observations within a child. The variables are (in column order):
ID = patient id
FEV1 = percent-predicted forced expiratory volume in 1 second AGE = age (years)
FEMALE = 1=female, 0=male
PSEUDOA = infection with Pseudo Aeruginosa (0=no, 3=yes)
F508 = genotype (1=homozygous, 2=heterozygous, 0=none)
PANCREAT = pancreatic enzyme supplmentation (0,1=no, 2=yes)
AGE0 = age at first follow up
AGEL = longitudinal component of age (AGE-AGE0)

Note: PSEUDOA and PANCREAT are just extra variables, which we are not going to use for this homework assignment.

Question 2

(a) Read 'dental-data.txt' file into SAS or R (depending upon your preference). The data do not have a header, so you will have to create variable names yourself.

Create 2 dummy variables from the variable genotype: F508_1 (1=if F508 is equal to '1', otherwise=0) and F508_2 (1=if F508 is equal to '2', otherwise=0)

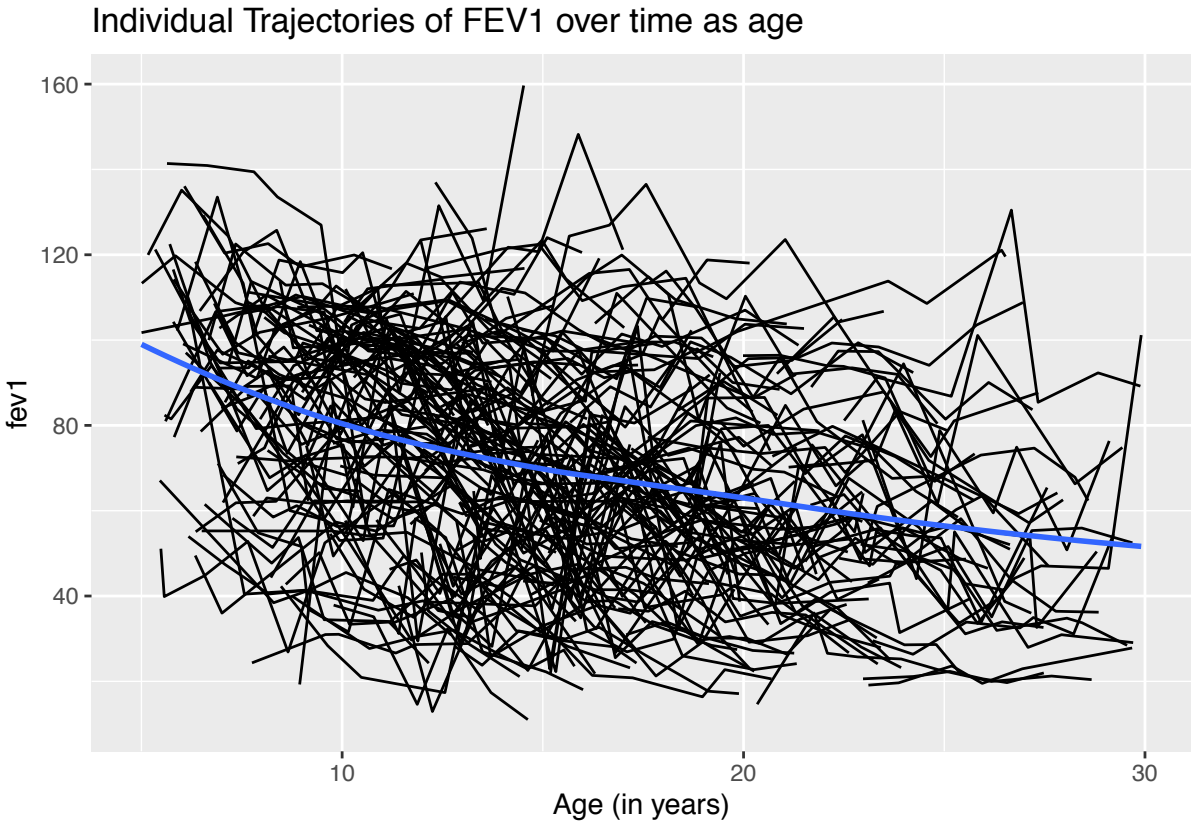
```
setDT(dt)
dt[,f508_1:=ifelse(f508==1,1,0)]
dt[,f508_2:=ifelse(f508==2,1,0)]
```

Create a new categorical variable for age at first follow up (*age0*): 1= if <8, 2= if >=8 and =<12, 3= if >12 and <15, 4= if >=15

```
dt[,agecat:=ifelse(age0 < 8, 1,
                   ifelse(age0 >= 8.000 & age0<=12, 2,
                          ifelse(age0 > 12 & age0 < 15, 3,
                                 ifelse(age0 >= 15, 4, NA))))]
```

(b) On a single graph, construct a time plot that displays individual trajectories of FEV1 with time as AGE, with a smoothed estimate of the population mean FEV1 over time. Describe what you see.

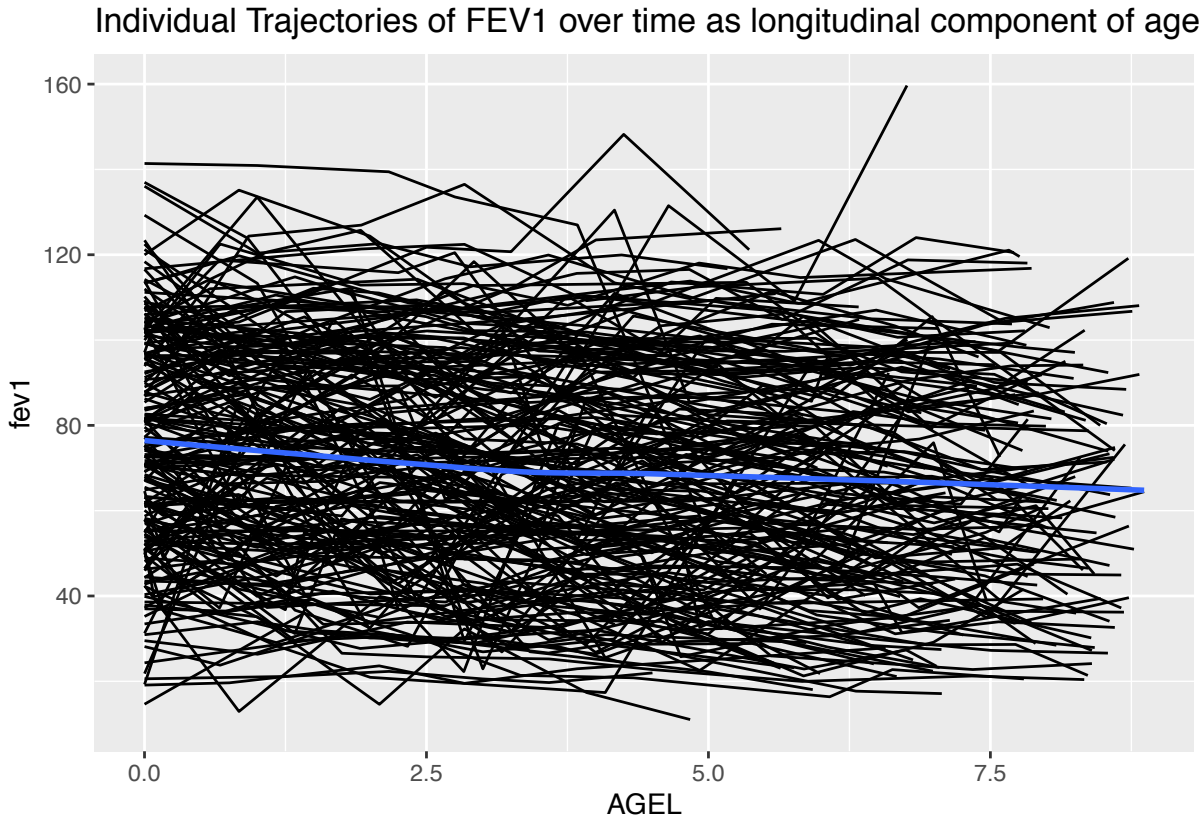
```
# Construct a time plot that displays individual trajectories of FEV1 with time as AGE
p1 <- ggplot(dt, aes(x=age, y=fev1))
p1 <- p1 + geom_line(aes(group=id)) +
  labs(title="Individual Trajectories of FEV1 over time as age", x="Age (in years)")

# With a smoothed estimate of the population mean FEV1 over time
p1 + geom_smooth(method='loess', se=F)
```



There's a trend that FEV1 decrease over the follow up time. The age is not equally spaced.

(c) On a single graph, construct a time plot that displays individual trajectories of FEV1 with time as AGEL, with a smoothed estimate of the population mean FEV1 over time. Describe what you see.

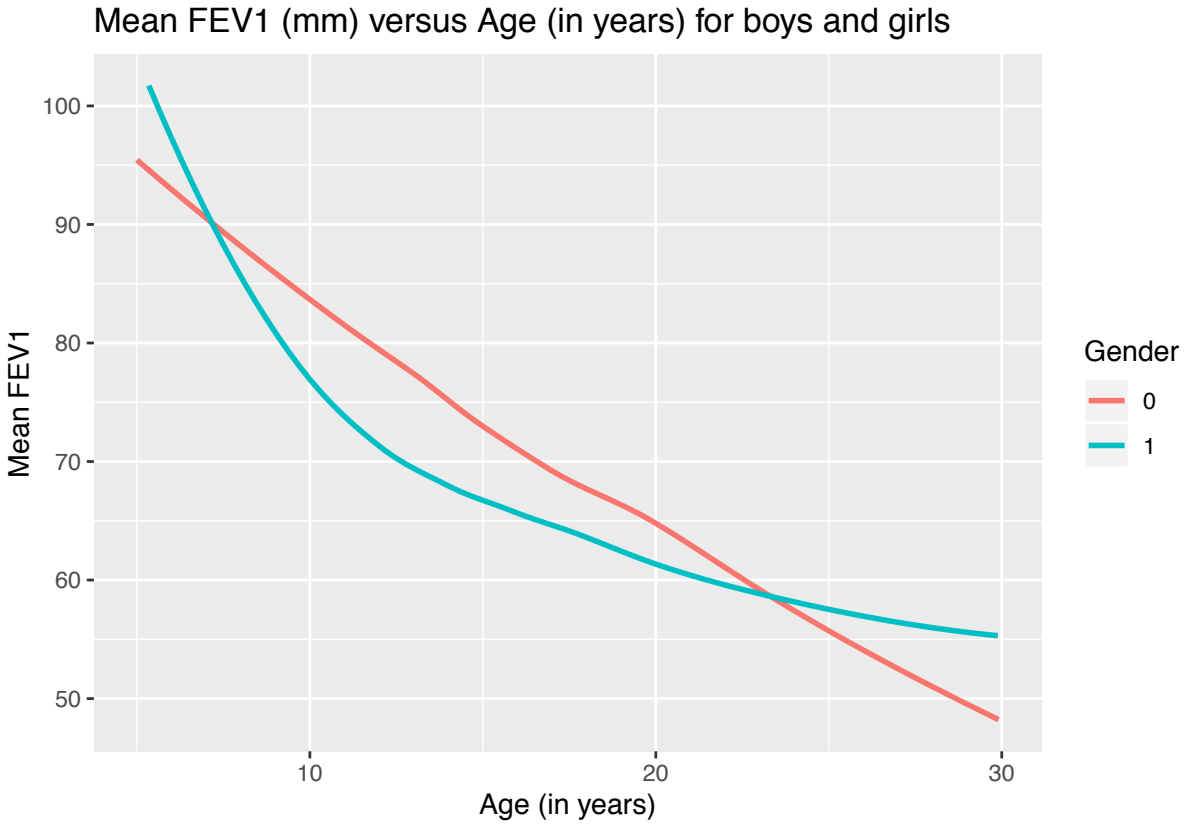
```
# Construct a time plot that displays individual trajectories of FEV1 with time as AGE
p2 <- ggplot(dt, aes(x=agel, y=fev1))
p2 <- p2 + geom_line(aes(group=id)) +
  labs(title="Individual Trajectories of FEV1 over time as longitudinal component of age (agel)", x="AG

# With a smoothed estimate of the population mean FEV1 over time
p2 + geom_smooth(method='loess', se=F)
```



FEV1 steadily decrease over the longitudinal time component.

(d) On a single graph, construct a time plot that displays mean FEV1 for boys and girls, with the time as AGE. Briefly describe the time trends for boys and girls.

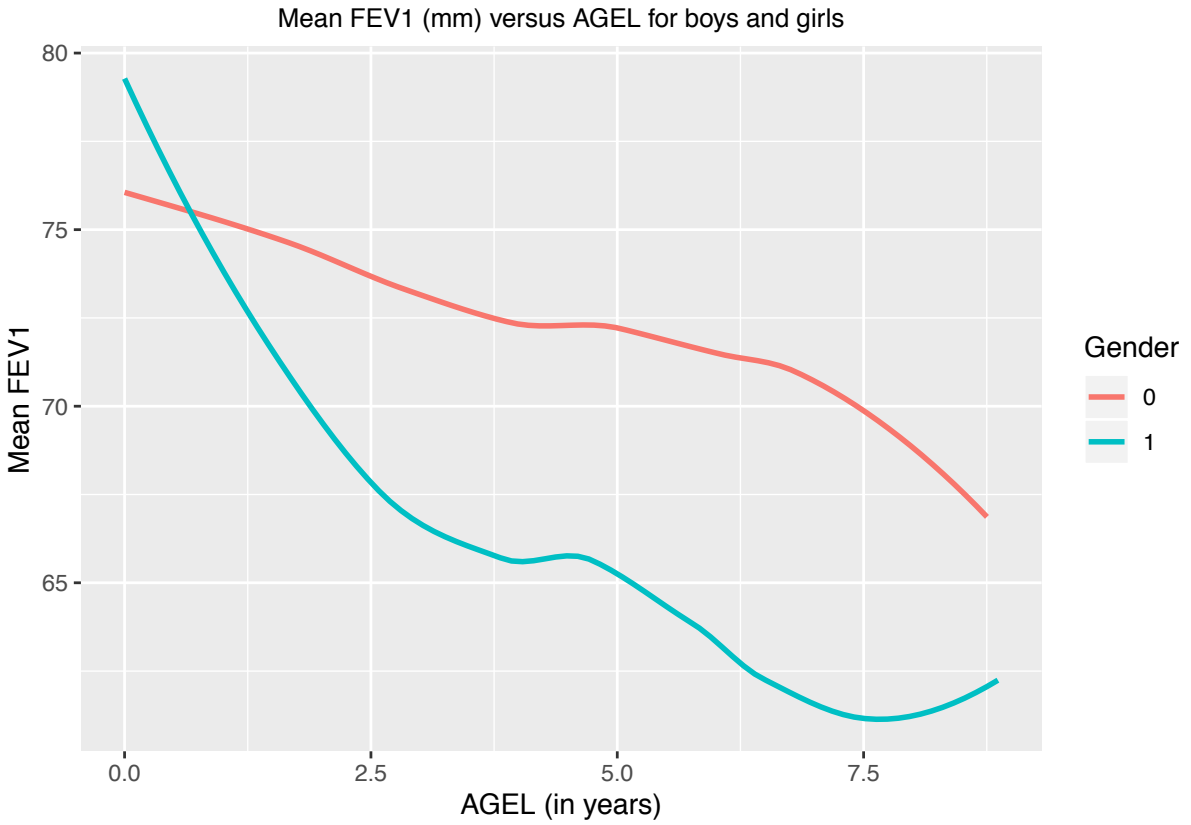
```
dt$female = factor(dt$female)

mean = dt %>%
  group_by(female, age) %>%
  summarize(mean_fev1 = mean(fev1))

p3 = ggplot(mean, aes(x = age, y = mean_fev1,group = female)) +
 # geom_line(aes(color=female)) +
  geom_smooth(aes(color=female),se=F,method="loess") +
  labs(title="Mean FEV1 (mm) versus Age (in years) for boys and girls", x="Age (in years)", y="Mean FEV
p3
```



Both boys and girls' mean FEV1 decrease over the age. Between ~age5-25, female's mean FEV1 are less than male's, while when age<5 or >25, female mean FEV1 is more than male's.

(e) On a single graph, construct a time plot that displays mean FEV1 for boys and girls, with the time as AGEL. Briefly describe the time trends for boys and girls.

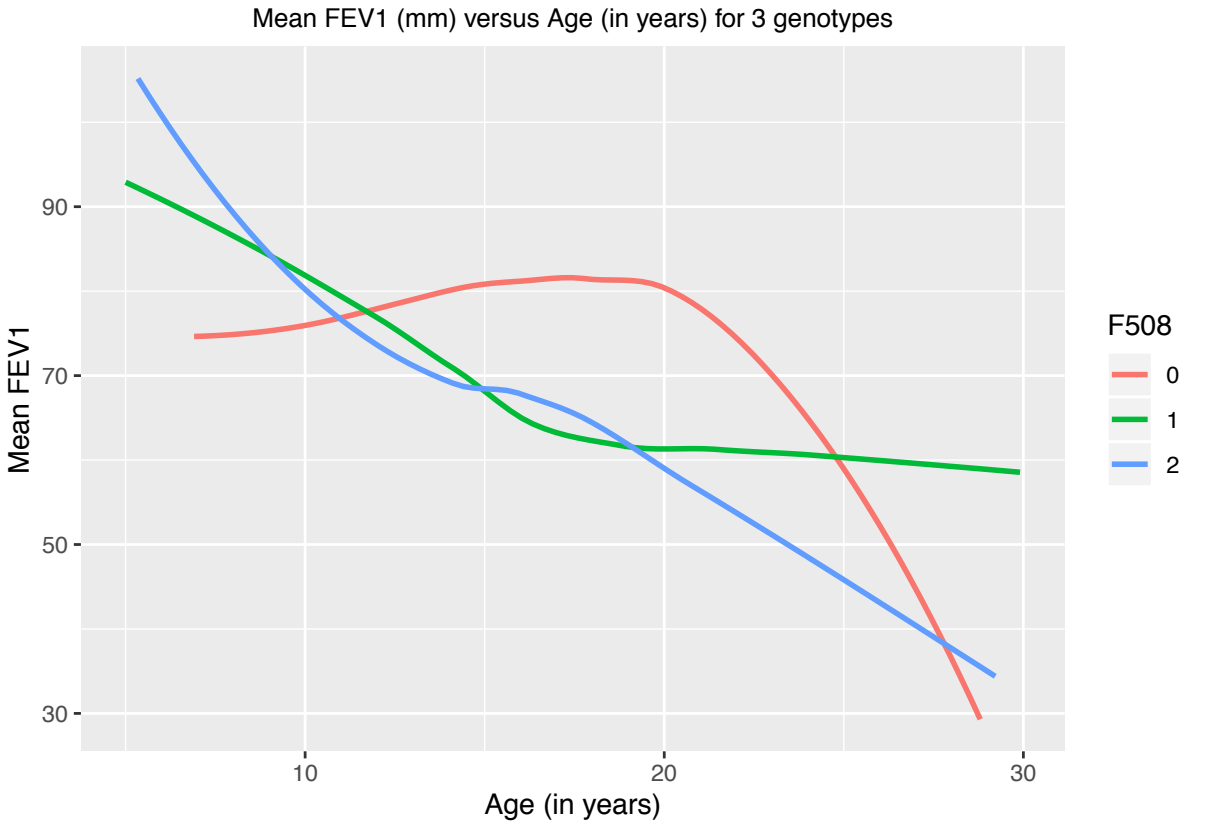
```
mean = dt %>%
  group_by(female, agel) %>%
  summarize(mean_fev1 = mean(fev1))
# mean

p4 = ggplot(mean, aes(x = agel, y = mean_fev1, group = female )) +
  geom_smooth(aes(color=female),se=F,method="loess") +
  labs(title="Mean FEV1 (mm) versus AGEL for boys and girls",
       x="AGEL (in years)", y="Mean FEV1", color = "Gender") +
  theme(plot.title = element_text(size = 10, hjust = 0.5))
p4
```



Both boys and girls' mean FEV1 decrease over the longitudinal age for the most of time. Female's mean FEV1 is less than male's when age $>$ ~0.5.

(f) On a single graph, construct a time plot that displays mean FEV1 by genotype, with the time as AGE. Briefly describe the time trends for the three genotype groups.

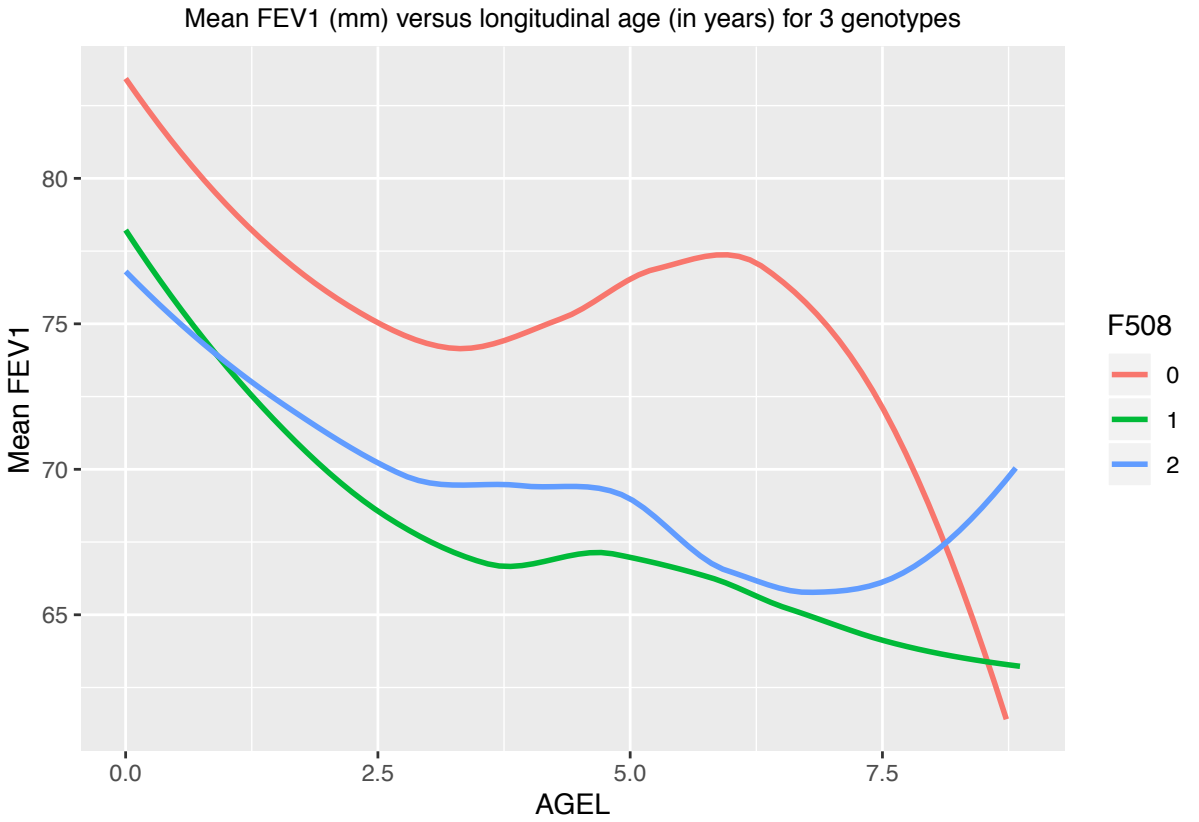
```
dt$f508 = factor(dt$f508)
mean = dt %>%
  group_by(f508, age) %>%
  summarize(mean_fev1 = mean(fev1))

p5 = ggplot(mean, aes(x = age, y = mean_fev1, group = f508)) +
 geom_smooth(aes(color=f508),se=F,method="loess") +
  labs(title="Mean FEV1 (mm) versus Age (in years) for 3 genotypes",
       x="Age (in years)", y="Mean FEV1", color = "F508") +
  theme(plot.title = element_text(size = 10, hjust = 0.5))
p5
```



Mean FEV1 for both homozygous f508 and heterozygous f508 groups decrease over age. For non-f508 group, the mean fev1 increase until age 20 and then decrease.

(g) On a single graph, construct a time plot that displays mean FEV1 by genotype, with the time as AGEL. Briefly describe the time trends for the three genotype groups.

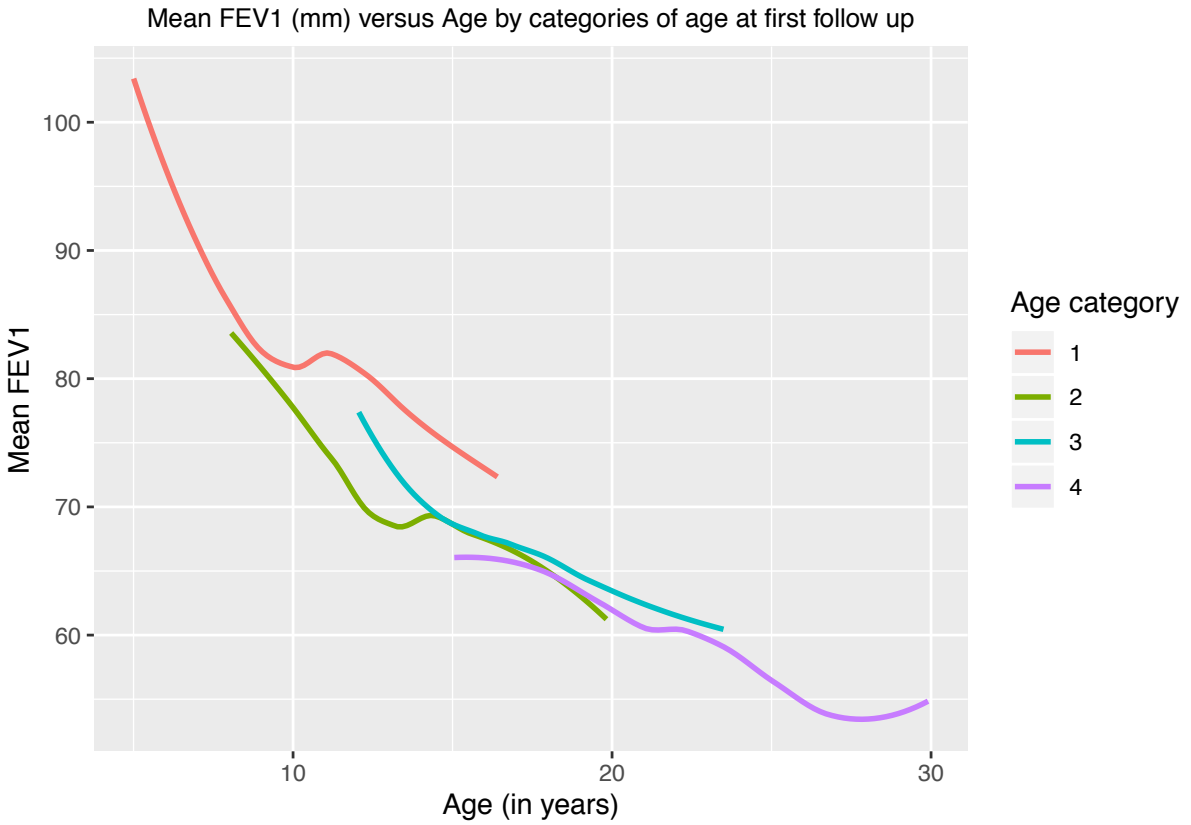
```
mean = dt %>%
  group_by(f508, agel) %>%
  summarize(mean_fev1 = mean(fev1))

p6 = ggplot(mean, aes(x = agel, y = mean_fev1,  group = f508)) +
  geom_smooth(aes(color=f508),se=F,method="loess") +
  labs(title="Mean FEV1 (mm) versus longitudinal age (in years) for 3 genotypes",
       x="AGEL", y="Mean FEV1", color = "F508") +
  theme(plot.title = element_text(size = 10, hjust = 0.5))
p6
```



Mean FEV1 for homozygous-f508 group decrease over agel, for heterozygous-f508 group decrease until agel = 7.5 and increase after, for non-f508 group decrease until agel=3.5 then increase until agel=6.25 then decrease again.

(h) On a single graph, construct a time plot that displays mean FEV1 by the new categorical variable of age at first follow up, with the time as AGE. Briefly describe the time trends for the four cohorts.

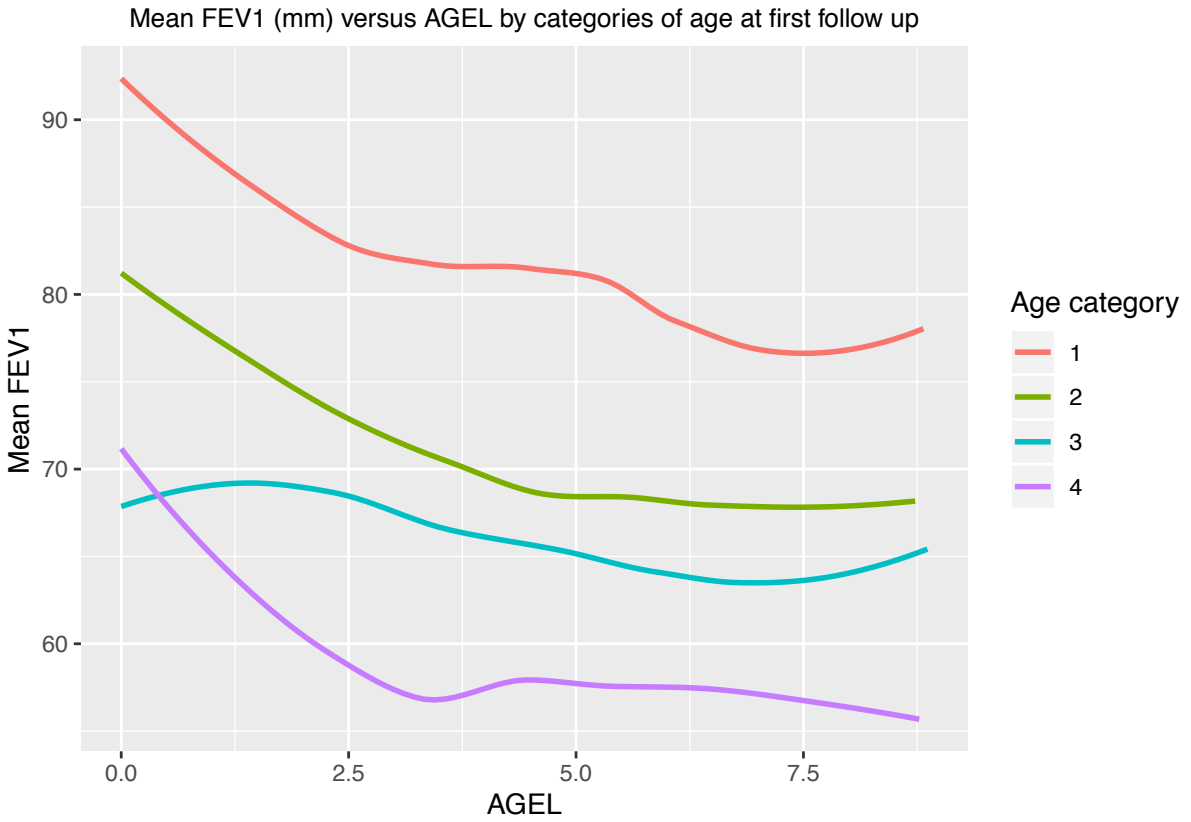
```
dt$agecat = factor(dt$agecat)
mean = dt %>%
  group_by(agecat, age) %>%
  summarize(mean_fev1 = mean(fev1))

p7 = ggplot(mean, aes(x = age, y = mean_fev1, group = agecat)) +
  geom_smooth(aes(color=agecat),se=F,method="loess") +
  labs(title="Mean FEV1 (mm) versus Age by categories of age at first follow up",
       x="Age (in years)", y="Mean FEV1", color = "Age category") +
  theme(plot.title = element_text(size = 10, hjust = 0.5))
p7
```



Mean FEV1 decrease over age for all four age category groups. The age<8-group has the most decrease rate from the graph.

(i) On a single graph, construct a time plot that displays mean FEV1 by the new categorical variable of age at first follow up, with the time as AGEL. Briefly describe the time trends for the four cohorts.

```
mean = dt %>%
  group_by(agecat, agel) %>%
  summarize(mean_fev1 = mean(fev1))

p8 = ggplot(mean, aes(x = agel, y = mean_fev1,  group = agecat )) +
  geom_smooth(aes(color=agecat),se=F,method="loess") +
  labs(title="Mean FEV1 (mm) versus AGEL by categories of age at first follow up",
       x="AGEL", y="Mean FEV1", color = "Age category") +
  theme(plot.title = element_text(size = 10, hjust = 0.5))
p8
```



Mean FEV1 decrease over agel for the most of four age category groups, except increase until agel=2.5 then decrease for the age in (12,15) group. The age<8-group has the most decrease rate from the graph.

Question 3: Consider the following maximal model: a fixed intercept (X1), age at first follow up (X2), longitudinal age (X3), female gender (X4), F508_1 (X5), F508_2 (X6), and 3 interactions: longitudinal age by female gender (X3 * X4), longitudinal age by F508_1 (X3 * X5), and longitudinal age by F508_2 (X3 * X6):

(a) **Fit a model for the random intercept only:**

```
###Fit a linear mixed effect model (LME) with random intercept to the data
library(nlme)
```

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

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

```
m1 <- lme(fev1 ~ age0 + agel + female + f508_1 + f508_2 + agel*female + agel*f508_1 + agel*f508_2, data=
          random= ~ 1 | id)
summary(m1)
```

```
## Linear mixed-effects model fit by REML
##  Data: dt
##        AIC      BIC    logLik
##   12520.08 12578.56 -6249.041
```

```
##
## Random effects:
##  Formula: ~1 | id
##         (Intercept) Residual
## StdDev:    22.59935 12.19734
##
## Fixed effects: fev1 ~ age0 + agel + female + f508_1 + f508_2 + agel * female +      agel * f508_1 + a
##                  Value Std.Error   DF   t-value p-value
## (Intercept)  103.80627  6.706026 1309 15.479550  0.0000
## age0          -1.85532  0.334312  195 -5.549663  0.0000
## agel          -0.58782  0.393060 1309 -1.495504  0.1350
## female1       -1.16203  3.397872  195 -0.341987  0.7327
## f508_1        -4.28096  5.611017  195 -0.762956  0.4464
## f508_2        -6.74044  5.642572  195 -1.194569  0.2337
## agel:female1  -0.82574  0.249427 1309 -3.310541  0.0010
## agel:f508_1   -0.48769  0.422469 1309 -1.154391  0.2486
## agel:f508_2   -0.65745  0.421359 1309 -1.560310  0.1189
##  Correlation:
##              (Intr) age0   agel   femal1 f508_1 f508_2 agl:f1 a:508_1
## age0         -0.630
## agel         -0.212  0.006
## female1      -0.164 -0.088  0.073
## f508_1       -0.657 -0.002  0.229 -0.021
## f508_2       -0.725  0.123  0.226 -0.062  0.788
## agel:female1  0.062 -0.005 -0.272 -0.262  0.004  0.011
## agel:f508_1   0.181 -0.004 -0.856  0.005 -0.267 -0.214 -0.022
## agel:f508_2   0.179 -0.003 -0.853  0.011 -0.215 -0.266 -0.041  0.805
##
## Standardized Within-Group Residuals:
##          Min           Q1          Med           Q3          Max
## -4.942954121 -0.528230691 -0.003676248  0.516817787  4.289855016
##
## Number of Observations: 1513
## Number of Groups: 200
```

i. Describe in words the interpretation of the random intercept component (b1i) and report the estimate of the variance for this random effect.

b_{1i}: Each subject has an underlying level of response which persists across all repeated measurements. The response of the ith subject at jth occasion differs from the population mean by a subject-level effect b_{1i}, and a within subject error, ϵ_{ij}. The variance is $22.59935^2 = 510.73$. Different subject has different b_{1i}.

```
getVarCov(m1)
```

```
## Random effects variance covariance matrix
##             (Intercept)
## (Intercept)      510.73
##   Standard Deviations: 22.599
```

ii. Output an estimated covariance matrix for Yi (sigma_i) for a child and describe the patterns in the variances and covariances. (Note: this covariance matrix is labeled 'Vi' in SAS.)

```r
library(mgcv)
a <- extract.lme.cov(m1)
a[1:7,1:7]
```

```
##          [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]
## [1,] 659.5058 510.7307 510.7307 510.7307 510.7307 510.7307 510.7307
## [2,] 510.7307 659.5058 510.7307 510.7307 510.7307 510.7307 510.7307
## [3,] 510.7307 510.7307 659.5058 510.7307 510.7307 510.7307 510.7307
## [4,] 510.7307 510.7307 510.7307 659.5058 510.7307 510.7307 510.7307
## [5,] 510.7307 510.7307 510.7307 510.7307 659.5058 510.7307 510.7307
## [6,] 510.7307 510.7307 510.7307 510.7307 510.7307 659.5058 510.7307
## [7,] 510.7307 510.7307 510.7307 510.7307 510.7307 510.7307 659.5058
```

It's in the format of compound symmetry, with constant variance across occasions, and constant correlation.

iii. Output empirical BLUPs and obtain their variance, then compare that variance to the estimate in Q3(a)(i). If they are different, why?

```r
m1.re <- random.effects(m1) #Obtain empirical BLUPs
head(m1.re,10)
```

```
##        (Intercept)
## 100073   24.273515
## 100111   24.932333
## 100185  -35.274079
## 100329  -11.038907
## 100352   15.085220
## 100636   20.122356
## 100736    4.400511
## 100815  -19.133568
## 100895   37.704748
## 100897   24.425543
```

```r
var(m1.re)
```

```
##             (Intercept)
## (Intercept)    481.4666
```

The variance of empirical BLUPS is smaller comparing with the variance in Q3(a)(i). Because the empirical BLUPs have been "shrunk" toward the population fixed effects, their distribution does not accurately represent the distribution of the random effects (e.g. due to "shrinkage" toward the population mean, empirical BLUPs have smaller variance).

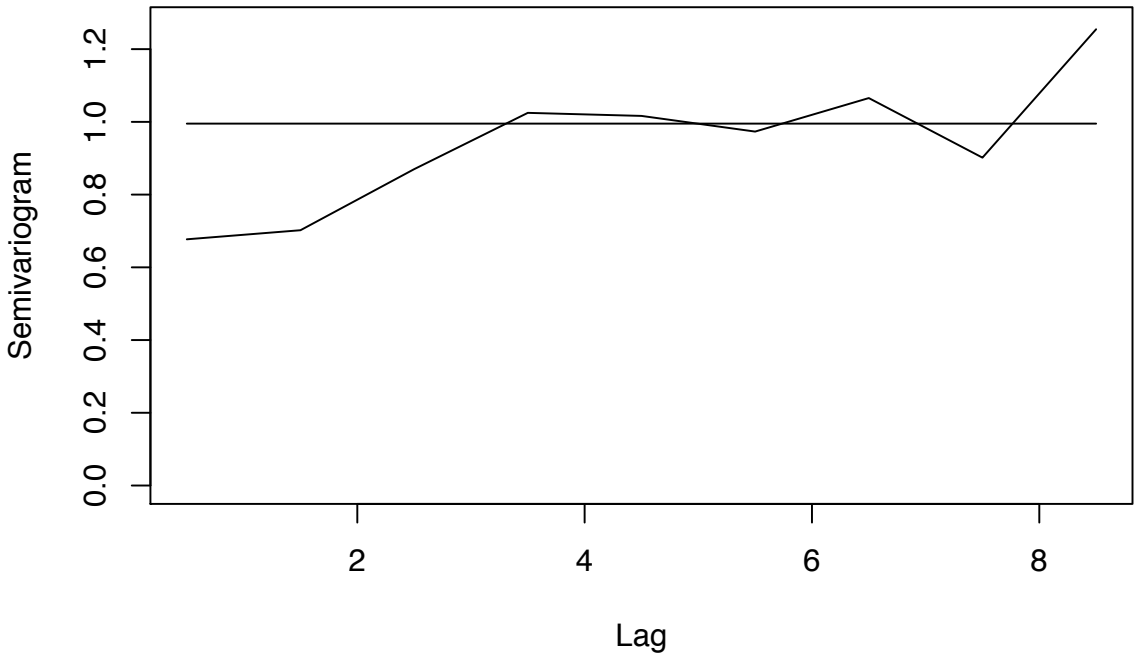
iv. Using Cholesky-transformed residuals,plot the empirical semivariogram and comment on whether the covariance model fits the data well or not.

```r
library(mgcv)
# Transform the residuals based on the Cholesky decomposition of the covariance matrix
est.cov <- extract.lme.cov(m1)
```

```
## Warning in model.matrix.default(~b$groups[[n.levels - i + 1]] - 1,
## contrasts.arg = c("contr.treatment", : non-list contrasts argument ignored
```

```r
cr <- solve(t(chol(est.cov))) %*% residuals(m1, level=0) #1:conditional, 0: marginal.

source("variogram.R")
```

```
#Empirical semi-variogram for transformed residuals
variogram(resid=cr[,1],timeVar=dt$agel,id=dt$id,irregular=T,binwidth=1,numElems=5)
```




```
## $bin.mids
## [1] 0.5 1.5 2.5 3.5 4.5 5.5 6.5 7.5 8.5
##
## $bin.means
## [1] 0.6770667 0.7019941 0.8699061 1.0248399 1.0163810 0.9733303 1.0654922
## [8] 0.9016683 1.2546967
##
## $bin.sizes
## [1]  586 1081  924  795  639  474  341  198   65
##
## $process.var
## [1] 0.995186
```

The curve is around 1, so I would say it fits data well.

(b) Fit a covariance pattern model with the Compound Symmetry for Ri;

```
m2 <- gls(fev1 ~ age0 + agel + female + f508_1 + f508_2 + agel:female + agel:f508_1 + agel:f508_2, data=
          corr=corCompSymm(form= ~ agel | id))
summary(m2)
```

```
## Generalized least squares fit by REML
##   Model: fev1 ~ age0 + agel + female + f508_1 + f508_2 + agel:female +      agel:f508_1 + agel:f508_
##   Data: dt
##        AIC      BIC    logLik
##   12520.08 12578.56 -6249.041
##
## Correlation Structure: Compound symmetry
##  Formula: ~agel | id
##  Parameter estimate(s):
```

```
##       Rho
## 0.7744143
##
## Coefficients:
##                  Value Std.Error   t-value p-value
## (Intercept)  103.80627  6.706025 15.479552  0.0000
## age0          -1.85532  0.334312 -5.549664  0.0000
## agel          -0.58782  0.393060 -1.495504  0.1350
## female1       -1.16203  3.397872 -0.341987  0.7324
## f508_1        -4.28096  5.611017 -0.762956  0.4456
## f508_2        -6.74044  5.642571 -1.194569  0.2324
## agel:female1  -0.82574  0.249427 -3.310541  0.0010
## agel:f508_1   -0.48769  0.422469 -1.154391  0.2485
## agel:f508_2   -0.65745  0.421359 -1.560310  0.1189
##
##  Correlation:
##               (Intr) age0   agel   femal1 f508_1 f508_2 agl:f1 a:508_1
## age0          -0.630
## agel          -0.212  0.006
## female1       -0.164 -0.088  0.073
## f508_1        -0.657 -0.002  0.229 -0.021
## f508_2        -0.725  0.123  0.226 -0.062  0.788
## agel:female1   0.062 -0.005 -0.272 -0.262  0.004  0.011
## agel:f508_1    0.181 -0.004 -0.856  0.005 -0.267 -0.214 -0.022
## agel:f508_2    0.179 -0.003 -0.853  0.011 -0.215 -0.266 -0.041  0.805
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.66032274 -0.73159979  0.01712711  0.73394138  3.18588347
##
## Residual standard error: 25.68084
## Degrees of freedom: 1513 total; 1504 residual
```

(c) When looking at the estimates of the fixed effects and standard errors from Q3(a) and Q3(b), what do you notice. Explain the results.

The estimates of the fixed effects and standard errors are the same for two models. This is because LME with random intercept is equivalent as CS model.

```
summary(m1)
```

```
## Linear mixed-effects model fit by REML
##  Data: dt
##        AIC      BIC    logLik
##   12520.08 12578.56 -6249.041
##
## Random effects:
##  Formula: ~1 | id
##         (Intercept) Residual
## StdDev:    22.59935 12.19734
##
## Fixed effects: fev1 ~ age0 + agel + female + f508_1 + f508_2 + agel * female +      agel * f508_1 + a
##                  Value Std.Error   DF   t-value p-value
## (Intercept)  103.80627  6.706026 1309 15.479550  0.0000
## age0          -1.85532  0.334312  195 -5.549663  0.0000
```

```
## agel          -0.58782  0.393060 1309 -1.495504  0.1350
## female1       -1.16203  3.397872  195 -0.341987  0.7327
## f508_1        -4.28096  5.611017  195 -0.762956  0.4464
## f508_2        -6.74044  5.642572  195 -1.194569  0.2337
## agel:female1  -0.82574  0.249427 1309 -3.310541  0.0010
## agel:f508_1   -0.48769  0.422469 1309 -1.154391  0.2486
## agel:f508_2   -0.65745  0.421359 1309 -1.560310  0.1189
##  Correlation:
##               (Intr) age0   agel   femal1 f508_1 f508_2 agl:f1 a:508_1
## age0          -0.630
## agel          -0.212  0.006
## female1       -0.164 -0.088  0.073
## f508_1        -0.657 -0.002  0.229 -0.021
## f508_2        -0.725  0.123  0.226 -0.062  0.788
## agel:female1   0.062 -0.005 -0.272 -0.262  0.004  0.011
## agel:f508_1    0.181 -0.004 -0.856  0.005 -0.267 -0.214 -0.022
## agel:f508_2    0.179 -0.003 -0.853  0.011 -0.215 -0.266 -0.041  0.805
##
## Standardized Within-Group Residuals:
##          Min           Q1          Med           Q3          Max
## -4.942954121 -0.528230691 -0.003676248  0.516817787  4.289855016
##
## Number of Observations: 1513
## Number of Groups: 200
```

```
summary(m2)
```

```
## Generalized least squares fit by REML
##   Model: fev1 ~ age0 + agel + female + f508_1 + f508_2 + agel:female +      agel:f508_1 + agel:f508_
##   Data: dt
##        AIC      BIC    logLik
##   12520.08 12578.56 -6249.041
##
## Correlation Structure: Compound symmetry
##  Formula: ~agel | id
##  Parameter estimate(s):
##       Rho
## 0.7744143
##
## Coefficients:
##                  Value Std.Error   t-value p-value
## (Intercept)  103.80627  6.706025 15.479552  0.0000
## age0          -1.85532  0.334312 -5.549664  0.0000
## agel          -0.58782  0.393060 -1.495504  0.1350
## female1       -1.16203  3.397872 -0.341987  0.7324
## f508_1        -4.28096  5.611017 -0.762956  0.4456
## f508_2        -6.74044  5.642571 -1.194569  0.2324
## agel:female1  -0.82574  0.249427 -3.310541  0.0010
## agel:f508_1   -0.48769  0.422469 -1.154391  0.2485
## agel:f508_2   -0.65745  0.421359 -1.560310  0.1189
##
##  Correlation:
##               (Intr) age0   agel   femal1 f508_1 f508_2 agl:f1 a:508_1
## age0          -0.630
## agel          -0.212  0.006
```

```
## female1      -0.164 -0.088  0.073
## f508_1       -0.657 -0.002  0.229 -0.021
## f508_2       -0.725  0.123  0.226 -0.062  0.788
## agel:female1  0.062 -0.005 -0.272 -0.262  0.004  0.011
## agel:f508_1   0.181 -0.004 -0.856  0.005 -0.267 -0.214 -0.022
## agel:f508_2   0.179 -0.003 -0.853  0.011 -0.215 -0.266 -0.041  0.805
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.66032274 -0.73159979  0.01712711  0.73394138  3.18588347
##
## Residual standard error: 25.68084
## Degrees of freedom: 1513 total; 1504 residual
```

Question 4: Consider the following maximal model (as in Q3): a fixed intercept (X1), age at first follow up (X2), longitudinal age (X3), female gender (X4), F508_1 (X5), F508_2 (X6), and 3 interactions: longitudinal age by female gender (X3 * X4), longitudinal age by F508_1 (X3 * X5), and longitudinal age by F508_2 (X3 * X6):

(a) Fit a model for the random intercept and random slope for the longitudinal age:

i. Describe in words the interpretation of the random intercept component (b1i), the random slope component (b3i); and Output the variance/covariance matrix for the random effects (G matrix) and its correlation matrix. Explain in words the meaning of the correlation estimate.

b_{1i}: Each subject has an underlying level of response which persists across all repeated measurements.
b_{3i}: Eacg subject has an underlying level of response which depends on the level X_3 which is longitudinal age.
The response of the ith subject at jth occasion differs from the population mean by subject-level effects, b_{1i} and b_{3i}, and a within subject error, ϵ_{ij}

```
m3 <- lme(fev1 ~ age0 + agel + female + f508_1 + f508_2 + agel:female + agel:f508_1 + agel:f508_2, data=
          random= ~ 1 +  agel | id)
summary(m3)
```

```
## Linear mixed-effects model fit by REML
##  Data: dt
##        AIC      BIC    logLik
##   12415.17 12484.28 -6194.586
##
## Random effects:
##  Formula: ~1 + agel | id
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev    Corr
## (Intercept) 22.636679 (Intr)
## agel         2.124407 -0.158
## Residual    10.863808
##
## Fixed effects: fev1 ~ age0 + agel + female + f508_1 + f508_2 + agel:female +      agel:f508_1 + agel
##                  Value Std.Error   DF   t-value p-value
## (Intercept)  104.51929  6.644334 1309 15.730589  0.0000
## age0          -1.91054  0.331047  195 -5.771212  0.0000
```

```
## agel          -0.60278  0.590295 1309 -1.021151  0.3074
## female1       -1.30048  3.370102  195 -0.385888  0.7000
## f508_1        -4.23809  5.563647  195 -0.761746  0.4471
## f508_2        -6.65233  5.594460  195 -1.189093  0.2358
## agel:female1  -0.76243  0.381213 1309 -2.000000  0.0457
## agel:f508_1   -0.50011  0.635751 1309 -0.786645  0.4316
## agel:f508_2   -0.74591  0.634490 1309 -1.175602  0.2400
##  Correlation:
##               (Intr) age0   agel   femal1 f508_1 f508_2 agl:f1 a:508_1
## age0          -0.630
## agel          -0.211  0.002
## female1       -0.164 -0.088  0.075
## f508_1        -0.657 -0.002  0.230 -0.021
## f508_2        -0.725  0.122  0.227 -0.062  0.788
## agel:female1   0.060 -0.003 -0.279 -0.265  0.005  0.012
## agel:f508_1    0.179 -0.001 -0.850  0.005 -0.269 -0.214 -0.023
## agel:f508_2    0.178  0.000 -0.845  0.012 -0.215 -0.268 -0.046  0.797
##
## Standardized Within-Group Residuals:
##          Min           Q1          Med           Q3          Max
## -5.328972545 -0.461786202  0.005172039  0.478384956  4.319667674
##
## Number of Observations: 1513
## Number of Groups: 200
```

```
getVarCov(m3)
```

```
## Random effects variance covariance matrix
##             (Intercept)    agel
## (Intercept)    512.4200 -7.6219
## agel            -7.6219  4.5131
##   Standard Deviations: 22.637 2.1244
```

```
cov2cor(getVarCov(m3))
```

```
## Random effects variance covariance matrix
##             (Intercept)     agel
## (Intercept)     1.00000 -0.15849
## agel           -0.15849  1.00000
##   Standard Deviations: 1 1
```

Since the correlation coefficient is negative, b_{1i} and b_{3i} are negatively correlated. One increase, another will decrease, vice versa.

ii. Output an estimated covariance matrix for Yi (Σi) for a child and describe the patterns in the variances and covariances. (Note: this covariance matrix is labeled 'Vi' in SAS.) How does this covariance matrix compare to the one in Q3(a)(ii)?

```
library(mgcv)
a <- extract.lme.cov(m3)
a[1:7,1:7]
```

```
##          [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]
## [1,] 630.4415 509.8964 502.2592 496.5199 482.8691 475.4225 466.9469
## [2,] 509.8964 625.8903 501.7277 497.1132 486.1379 480.1507 473.3363
## [3,] 502.2592 501.7277 618.1408 498.9093 496.0330 494.4640 492.6782
```

```
## [4,] 496.5199 497.1132 498.9093 618.2813 503.4692 505.2204 507.2136
## [5,] 482.8691 486.1379 496.0330 503.4692 639.1784 530.8044 541.7858
## [6,] 475.4225 480.1507 494.4640 505.2204 530.8044 662.7828 560.6451
## [7,] 466.9469 473.3363 492.6782 507.2136 541.7858 560.6451 700.1327
```

It no longer has the pattern of CS.

iii. Output empirical BLUPs and obtain their variances, then compare these variances to the estimates in Q4(a)(i). If there are differences, why?

```
m3.re <- random.effects(m3) #Obtain empirical BLUPs
head(m3.re,10)
```

```
##        (Intercept)       agel
## 100073   23.877821  0.1893093
## 100111   20.251959  1.5724037
## 100185  -31.043340 -1.6208560
## 100329   -8.844190 -0.6821591
## 100352   17.520146 -0.5951109
## 100636   21.512249 -0.4733118
## 100736    1.571885  1.1429671
## 100815  -15.000483 -1.2709653
## 100895   38.920844 -0.4256432
## 100897   22.895756  0.5854052
```

```
var(m3.re)
```

```
##             (Intercept)      agel
## (Intercept)  467.596013 -1.934672
## agel          -1.934672  2.786418
```

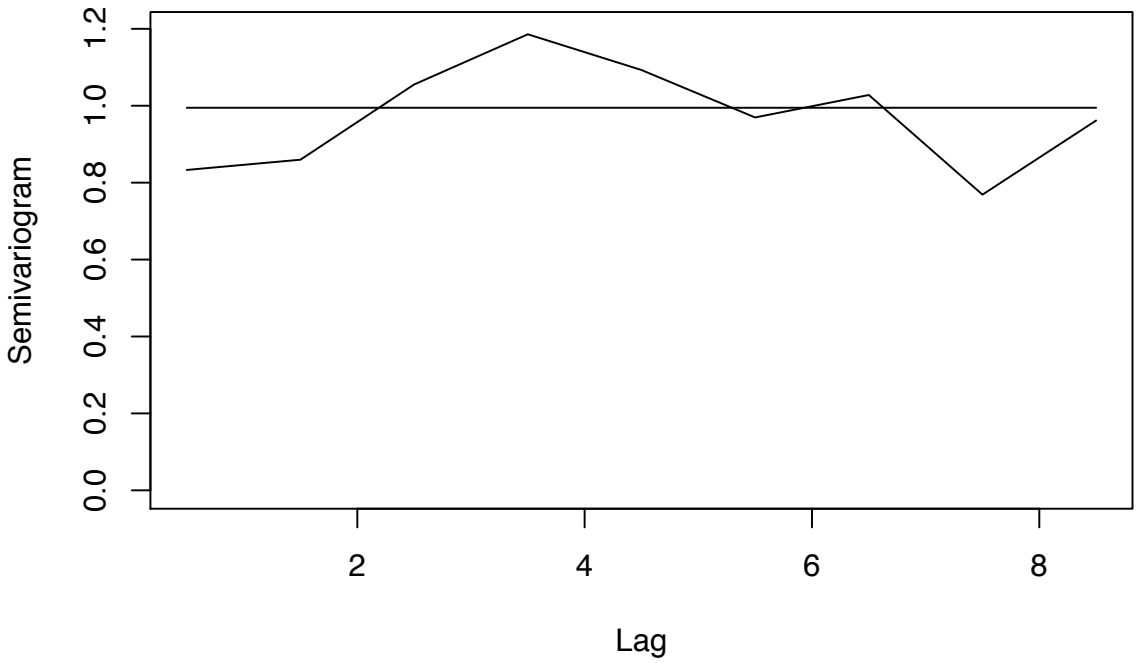
The variance of empirical BLUPS is smaller comparing with the variance in Q4(a)(i). Because the empirical BLUPs have been "shrunk" toward the population fixed effects, their distribution does not accurately represent the distribution of the random effects (e.g. due to "shrinkage" toward the population mean, empirical BLUPs have smaller variance).

iv. Using Cholesky-transformed residuals,plot the empirical semivariogram and comment on whether the covariance model fits the data well or not.

```
#Create a data frame with variables needed for residual analysis
library(mgcv)
# Transform the residuals based on the Cholesky decomposition of the covariance matrix
est.cov <- extract.lme.cov(m3)
```

```
## Warning in model.matrix.default(~b$groups[[n.levels - i + 1]] - 1,
## contrasts.arg = c("contr.treatment", : non-list contrasts argument ignored
```

```
cr <- solve(t(chol(est.cov))) %*% residuals(m3, level=0)
#pred4 <- data.frame(id=cd42$id,ctrt=cd42$ctrt,week=cd42$week,pred=fitted(m1, level=0), resid=cr[,1])

source("variogram.R")
#Empirical semi-variogram for transformed residuals
variogram(resid=cr[,1],timeVar=dt$agel,id=dt$id,irregular=T,binwidth=1,numElems=5)
```




```
## $bin.mids
## [1] 0.5 1.5 2.5 3.5 4.5 5.5 6.5 7.5 8.5
##
## $bin.means
## [1] 0.8330683 0.8596459 1.0552555 1.1858065 1.0929488 0.9695348 1.0279039
## [8] 0.7688627 0.9615472
##
## $bin.sizes
## [1]  586 1081  924  795  639  474  341  198   65
##
## $process.var
## [1] 0.9947167
```

It fits the data well because it's around 1.

(b) Compare the covariance model that uses only a random intercept (from Q3) to the covariance model in Q4 (random intercept and slope). Conduct appropriate hypothesis test and show your work (e.g., which parameters are being tested and what test are you using?). Which covariance model do you choose?

We want to test if the random slope which is the coefficiant, b_3, is statistically significant, i e. $H_0 : b_3 = 0, H_1 : b_3 \neq 0$. Since our two models are nested, we conduct a likelihood ratio test.

```
anova(m1,m3)
```

```
##    Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## m1     1 11 12520.08 12578.56 -6249.041
## m3     2 13 12415.17 12484.28 -6194.586 1 vs 2 108.9091  <.0001
```

The result gives p<0.0001 which is statistically significant. Therefore, we choose the model with the maximum log likelihood which is m3 which is with the random slope one.

Question 5: Starting with the "maximal" model for the mean response as in Q3 and Q4, and the best covariance model using random effect(s) specification (from Q4(b)), refine the model for the mean response (use a two-sided alpha of 0.05 for the hypotheses tests):

(a) Test whether having two fixed effects for age (cross-sectional and longitudinal) is needed, or whether just a cohort-averaged age effect is sufficient. What are the null and alternative hypotheses? Conduct your test. Explain your finding in words.

$$H_0 : \beta_{age0} = \beta_{agel}, H_1 : \beta_{age0} \neq \beta_{agel}$$

```
# Note to myself: put on the mean level - use non-factor. put on the ref level- use factor
m4 <- lme(fev1 ~ age0 + agel + female + f508_1 + f508_2, data=dt,
          random= ~ agel | id)
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

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

```
linearHypothesis(m4, c("agel = age0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## - age0  + agel = 0
##
## Model 1: restricted model
## Model 2: fev1 ~ age0 + agel + female + f508_1 + f508_2
##
##   Df  Chisq Pr(>Chisq)
## 1
## 2  1 1.0138      0.314
```

Since p = 0.314 > 0.05, we fail to reject the null hypothesis and conclude that having two fixed effects for age (cross-sectional and longitudinal) is not needed. They don't have statistically significant difference.

(b) After deciding whether to keep two fixed effects for age (variables AGE0 and AGEL) or only one fixed effect for age (variable AGE), test whether there is a significant contribution of genotype to the change in FEV1 over time. Explain your finding in words.

age = age0+agel. We decide to use age.

```
m5 <- lme(fev1 ~ age + female + f508_1 + f508_2 + age*f508_1 + age*f508_2, data=dt,
          random= ~ age | id)
anova(m5, type = "marginal")
```

```
##             numDF denDF   F-value p-value
## (Intercept)     1  1310 102.14697  <.0001
## age             1  1310   2.86809  0.0906
## female          1   196   1.47168  0.2265
## f508_1          1   196   0.02614  0.8717
```

```
## f508_2          1   196   0.24337  0.6223
## age:f508_1      1  1310   1.04944  0.3058
## age:f508_2      1  1310   2.55550  0.1102
```

Wald test results shows that the p value of two interaction terms are > 0.05, therefore, we fail to reject the null and conclude that there is not a significant contribution of genotype to the change in FEV1 over time.

(c) After deciding whether to keep two fixed effects for age (variables AGE0 and AGEL) or only one fixed effect for age (variable AGE), test whether there is a significant contribution of gender to the change in FEV1 over time. Explain your finding in words.

```
m5 <- lme(fev1 ~ f508_1 + f508_2 + age*female, data=dt,
          random= ~ age | id)
anova(m5, type = "marginal")
```

```
##             numDF denDF   F-value p-value
## (Intercept)     1  1311 250.04432  <.0001
## f508_1          1   196   1.51819  0.2194
## f508_2          1   196   2.15042  0.1441
## age             1  1311  27.93019  <.0001
## female          1   196   0.49979  0.4804
## age:female      1  1311   2.96145  0.0855
```

Again, Wald test results shows that the p value of the interaction terms $= 0.0855 > 0.05$, therefore, we fail to reject the null and conclude that there is not a significant contribution of gender to the change in FEV1 over time.

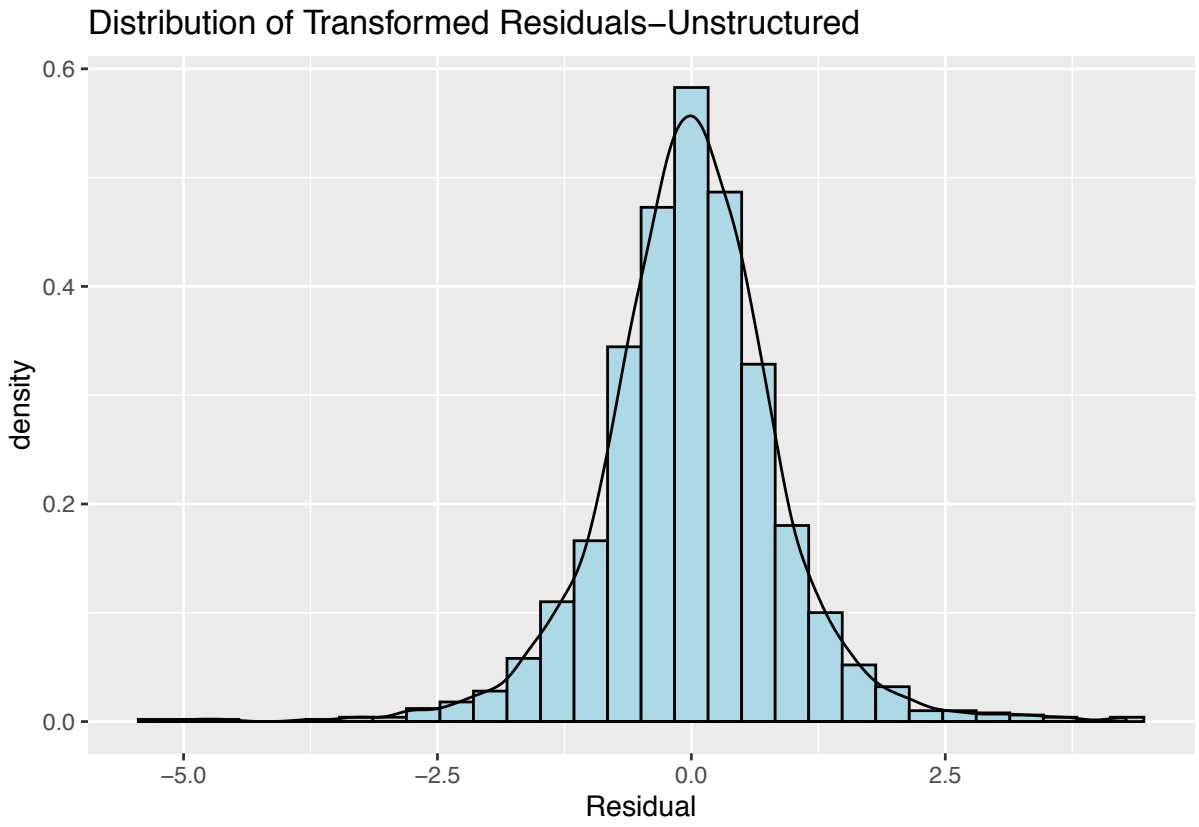
(d) After going through Q5(a,b,c), you should have built a model for the mean response that includes an intercept, an effect of age (either one variable or two variables), as well as the main effects of gender and genotype (keep these two variables in the model because we want to have the mean responses adjusted for these variables), and significant interaction effects from Q5(b,c):

i. Conduct appropriate residual diagnostics for the mean response model.

- Histogram of the transformed residuals

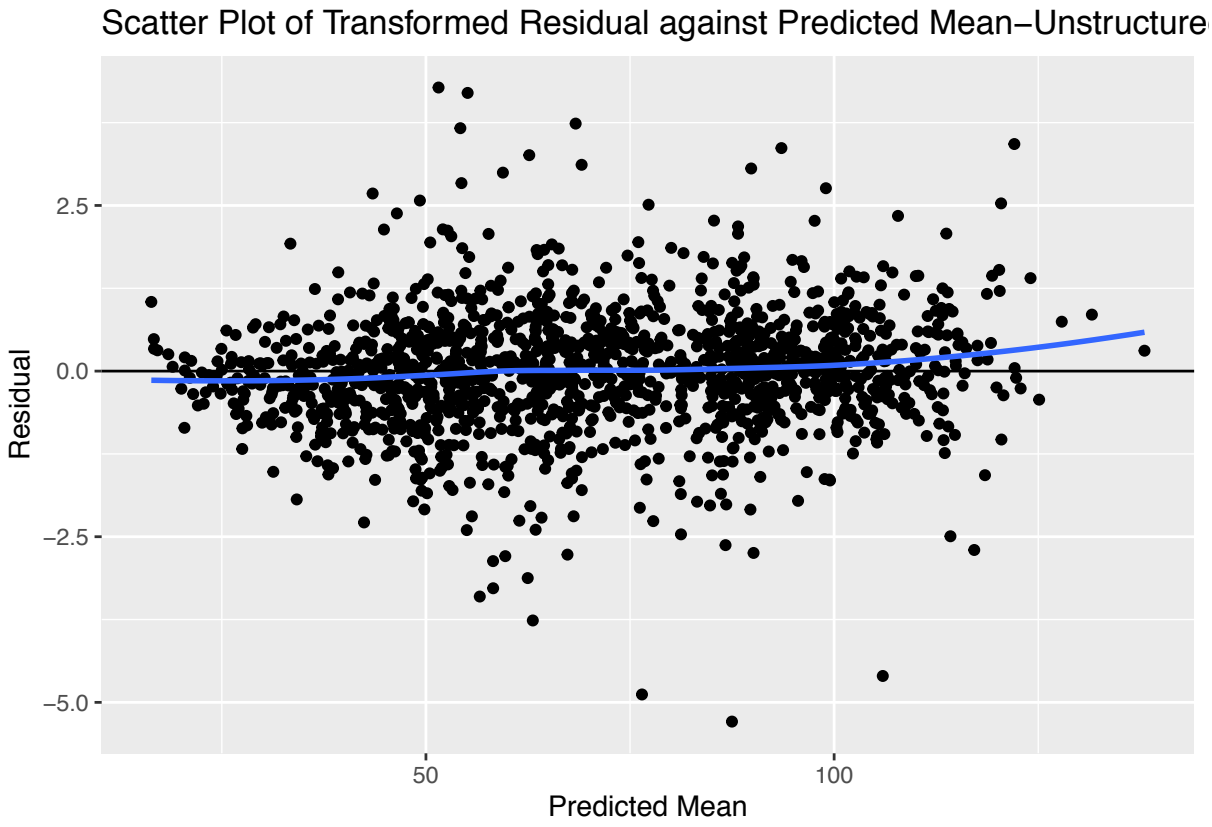
```
m6 <- lme(fev1 ~ age + female + f508_1 + f508_2, data=dt,
          random= ~ age | id)

pred1 <- data.frame(id=dt$id,gender=dt$female,age=dt$age,
                    pred=fitted(m6), resid=residuals(m6, type="normalized"))

p11 <- ggplot(pred1, aes(x=resid))
p11 + geom_histogram(aes(y=..density..), color="black", fill="lightblue") +
  geom_density() + labs(title="Distribution of Transformed Residuals-Unstructured", x="Residual")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



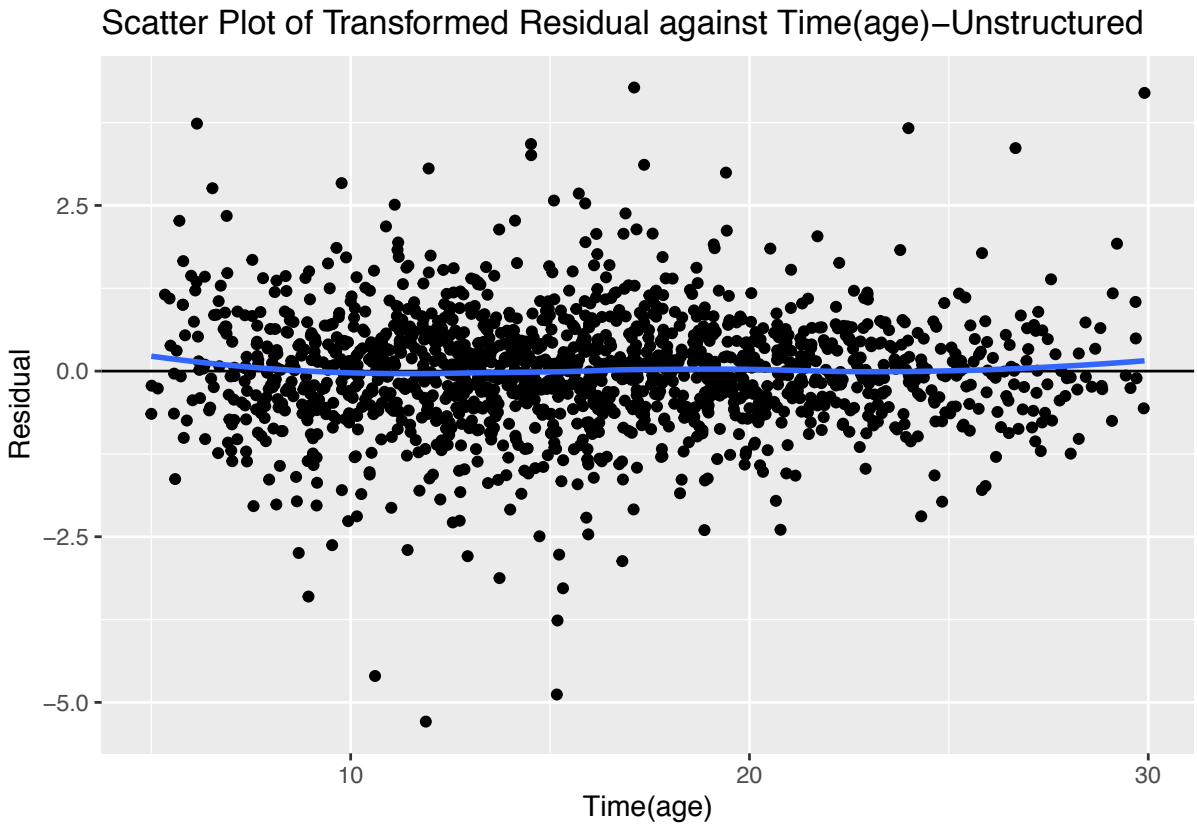
- Transformed residuals by predicted marginal mean response

```
p13 <- ggplot(pred1, aes(x=pred, y=resid))
p13 + geom_hline(yintercept=0,colour="black") +
  geom_point() +  geom_smooth(method="loess", se=F) +
  labs(title="Scatter Plot of Transformed Residual against Predicted Mean-Unstructured",
       x="Predicted Mean", y="Residual")
```



- Transformed residuals by the effect of time (AGE or AGEL, depending upon your answer to Q5(a))

```
p14 <- ggplot(pred1, aes(x=age, y=resid))
p14 + geom_hline(yintercept=0,colour="black") +
  geom_point() +  geom_smooth(method="loess", se=F) +
  labs(title="Scatter Plot of Transformed Residual against Time(age)-Unstructured",
       x="Time(age)", y="Residual", color="Treatment Group")
```



(e) Describe your final model for the mean response (mean FEV1) over time and any significant contributors (fixed effects), as well as the covariance model (random effects). The answer to Q5(e) should be summarized in 2-3 sentences and is the conclusion for your analyses – imagine that you are writing a conclusion in an abstract.

```
summary(m6)
```

```
## Linear mixed-effects model fit by REML
##  Data: dt
##        AIC      BIC    logLik
##   12428.57 12476.43 -6205.283
##
## Random effects:
##  Formula: ~age | id
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev   Corr
## (Intercept) 34.08536 (Intr)
## age          1.94745 -0.756
## Residual    10.96972
##
## Fixed effects: fev1 ~ age + female + f508_1 + f508_2
##                 Value Std.Error   DF   t-value p-value
## (Intercept) 103.74334  5.758901 1312 18.014435  0.0000
## age          -1.58044  0.176512 1312 -8.953735  0.0000
## female1      -4.27265  3.344957  196 -1.277340  0.2030
## f508_1       -7.17903  5.564796  196 -1.290079  0.1985
## f508_2       -8.51242  5.522549  196 -1.541393  0.1248
```

```
##  Correlation:
##         (Intr) age    femal1 f508_1
## age     -0.442
## female1 -0.256 -0.017
## f508_1  -0.762  0.008 -0.011
## f508_2  -0.770  0.030 -0.038  0.794
##
## Standardized Within-Group Residuals:
##          Min           Q1          Med           Q3          Max
## -5.290457581 -0.492517068  0.005302403  0.475077028  4.280264059
##
## Number of Observations: 1513
## Number of Groups: 200
```

Our final model is

$$Y_{ij} = \beta_1 + \beta_2 age_{ij} + \beta_3 \text{gender}_i + \beta_4 F5081_{ij} + \beta_5 F5082_{ij} + b_{1i} + b_{2i} age_{ij} + \epsilon_{ij}.$$

Mean FEV1 decrease over the age. A single year increase in age is associated 1.58 decrease in FEV1. Female has less mean FEV1 than male and people with F508 or/and F509 has less mean FEV1 than people without. Age is significant contributors to the fixed effects and contributes to the random effect in our model as well.