Statistics and R short course

Marc Henrion

18 October 2019

Session 6 - Practical (Solutions)

Exercise 1

- Return to the `Orthodont` dataset. We did not specify explicitly whether to use ML or REML estimation. Rerun the LMM we fitted for this dataset, but now with ML rather than REML estimation.
- Return to the `sleepstudy` dataset. Find out how you would specify different covariance structures and try out models with different structures.

Exercise 1 (Solution)

`Orthodont` dataset

Let's first use the `nlme` package

Recall from lectures:

```
# nlme package - REML
orthModLme_REML<-lme(distance~age+Sex,
                random=~1|Subject,
                data=Orthodont)
summary(orthModLme_REML)
## Linear mixed-effects model fit by REML
##  Data: Orthodont
##        AIC      BIC    logLik
##   447.5125 460.7823 -218.7563
##
## Random effects:
##  Formula: ~1 | Subject
##         (Intercept) Residual
## StdDev:    1.807425 1.431592
##
## Fixed effects: distance ~ age + Sex
##                 Value Std.Error DF   t-value p-value
## (Intercept) 17.706713 0.8339225 80 21.233044  0.0000
## age          0.660185 0.0616059 80 10.716263  0.0000
## SexFemale   -2.321023 0.7614168 25 -3.048294  0.0054
##  Correlation:
##           (Intr) age
## age       -0.813
## SexFemale -0.372  0.000
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -3.74889609 -0.55034466 -0.02516628  0.45341781  3.65746539
```

```
##
## Number of Observations: 108
## Number of Groups: 27
```

Now, changing from restricted maximum likelihood estimation to maximum likelihood estimation:

```
# nlme package - REML
orthModLme_ML<-lme(distance~age+Sex,
                random=~1|Subject,
                data=Orthodont,
                method="ML")
summary(orthModLme_ML)
## Linear mixed-effects model fit by maximum likelihood
##  Data: Orthodont
##        AIC      BIC    logLik
##   444.8565 458.2671 -217.4282
##
## Random effects:
##  Formula: ~1 | Subject
##         (Intercept) Residual
## StdDev:    1.730079 1.422728
##
## Fixed effects: distance ~ age + Sex
##                 Value Std.Error DF   t-value p-value
## (Intercept) 17.706713 0.8315459 80 21.293729  0.0000
## age          0.660185 0.0620929 80 10.632212  0.0000
## SexFemale   -2.321023 0.7430668 25 -3.123572  0.0045
##  Correlation:
##             (Intr) age
## age         -0.821
## SexFemale -0.364  0.000
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -3.77682007 -0.55426744 -0.01578248  0.45835495  3.68124620
##
## Number of Observations: 108
## Number of Groups: 27
```

We note that:

- The fixed effects are unchanged (this is not generally true; the fixed effects estimates are different between ML and REML, but this difference is usually small).
- The fixed effects standard errors are different (biased downwards in both ML and REML, but in ML this is compounded by the bias in the covariance parameters).
- The biggest change is observed in the estimate of the random effects variance parameter: the reported standard deviation is 1.730 (ML) vs. 1.807 (REML).

We observe the same thing, when using the `lme4` package:

```
# lme4 package - REML
orthModLmer_REML<-lmer(distance~age+Sex+(1|Subject),
                    data=Orthodont)
summary(orthModLmer_REML)
## Linear mixed model fit by REML ['lmerMod']
## Formula: distance ~ age + Sex + (1 | Subject)
```

```
##    Data: Orthodont
##
## REML criterion at convergence: 437.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.7489 -0.5503 -0.0252  0.4534  3.6575
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Subject  (Intercept) 3.267    1.807
##  Residual             2.049    1.432
## Number of obs: 108, groups:  Subject, 27
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 17.70671    0.83392  21.233
## age          0.66019    0.06161  10.716
## SexFemale   -2.32102    0.76142  -3.048
##
## Correlation of Fixed Effects:
##           (Intr) age
## age       -0.813
## SexFemale -0.372  0.000

# lme4 package - ML
orthModLmer_ML<-lmer(distance~age+Sex+(1|Subject),
                     data=Orthodont,
                     REML=FALSE)
summary(orthModLmer_ML)
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: distance ~ age + Sex + (1 | Subject)
##    Data: Orthodont
##
##      AIC      BIC   logLik deviance df.resid
##    444.9    458.3   -217.4    434.9      103
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.7768 -0.5543 -0.0158  0.4584  3.6812
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Subject  (Intercept) 2.993    1.730
##  Residual             2.024    1.423
## Number of obs: 108, groups:  Subject, 27
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 17.70671    0.81992  21.596
## age          0.66019    0.06122  10.783
## SexFemale   -2.32102    0.73267  -3.168
##
```

```
## Correlation of Fixed Effects:
##            (Intr) age
## age        -0.821
## SexFemale -0.364  0.000
```

`lme4` reports the variance rather than the standard deviation of the random effects. This is 2.993 (ML) vs. 3.267 (REML).

`sleepstudy` dataset

Recall from lectures:

```
# nlme package
sleepModLme<-lme(Reaction~Days,
                 random=~Days|Subject,
                 data=sleepstudy)
summary(sleepModLme)

# lme4 package
sleepModLmer<-lmer(Reaction~Days+(Days|Subject),
                   data=sleepstudy)
summary(sleepModLmer)
```

There are 2 covariance matrices for which we can attempt to specify different structures:

- The $\mathbf{D}$ matrix - the covariance matrix between random effects (in this case the random intercept and the random slope for the subject random factor).
- The $\mathbf{R}_i$ matrix - the within-group covariance structure.

The former is specified by the `random` argument in the `lme()` function call and the latter by specifying the argument `correlation`.

By default, $\mathbf{D}$ is unstructured (i.e. a covariance parameter for every pair of random effects) and $\mathbf{R}_i$ is a diagonal matrix.

Specifying the D matrix

We wish to specify a diagonal $\mathbf{D}$ matrix rather than the unstructured variance-covariance matrix that is used by default. How we do this, differs between the packages.

To specify independent random effects, using `lme()` from the `nlme` package we need to specify distinct random effects:

```
# nlme package
sleepModLmeDiagD<-lme(Reaction~Days,
                 random=list(~1|Subject, # random intercept
                             ~0+Days|Subject), # independent random slope
                 data=sleepstudy)
summary(sleepModLmeDiagD)
## Linear mixed-effects model fit by REML
##  Data: sleepstudy
##        AIC      BIC    logLik
##   1753.669 1769.578 -871.8346
##
## Random effects:
##  Formula: ~1 | Subject
```

```
##         (Intercept)
## StdDev:    25.05133
##
##  Formula: ~0 + Days | Subject %in% Subject
##             Days Residual
## StdDev: 5.988172 25.56529
##
## Fixed effects: Reaction ~ Days
##                 Value Std.Error  DF  t-value p-value
## (Intercept) 251.40510  6.885381 161 36.51288       0
## Days         10.46729  1.559566 161  6.71167       0
##  Correlation:
##      (Intr)
## Days -0.184
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -3.96258544 -0.46251664  0.02039796  0.46525704  5.18601943
##
## Number of Observations: 180
## Number of Groups:
##                Subject Subject.1 %in% Subject
##                     18                    18
```

Using `lmer()` from the `lme4` package, we can use the double bar notation: `||`:

```
# lme4 package
sleepModLmerDiagD<-lmer(Reaction~Days+(Days||Subject),
                  data=sleepstudy)
summary(sleepModLmerDiagD)
## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction ~ Days + ((1 | Subject) + (0 + Days | Subject))
##    Data: sleepstudy
##
## REML criterion at convergence: 1743.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9626 -0.4626  0.0204  0.4653  5.1860
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  Subject   (Intercept) 627.50   25.050
##  Subject.1 Days         35.86    5.989
##  Residual              653.58   25.565
## Number of obs: 180, groups:  Subject, 18
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  251.405      6.885  36.514
## Days          10.467      1.560   6.711
##
## Correlation of Fixed Effects:
##      (Intr)
```

```
## Days -0.184
```

Specifying the $\mathbf{R}_i$ matrix

We cannot change the default structure (diagonal) that is used by `lmer()` in the `lme4` package. Contrary to `nlme`, when using `lme4`, the correlation structure between residuals can only be modelled through random effects. If we wish to specify an alternative structure, we have to use `lme()` from the `nlme` package.

Let's try to fit 2 models with different structures for the $\mathbf{R}_i$:

- diagonal, but heterogeneous (different variances for different individuals)
- compound symmetry

```
# nlme package

# heterogeneous, diagonal
sleepModLmeHetDiagR<-lme(Reaction~Days,
                 random=~Days|Subject,
                 weights=varIdent(form=~1|Subject), # heterogeneity accounted for through weights matri
                 data=sleepstudy)
summary(sleepModLmeHetDiagR)
## Linear mixed-effects model fit by REML
##  Data: sleepstudy
##        AIC      BIC    logLik
##   1712.251 1785.432 -833.1256
##
## Random effects:
##  Formula: ~Days | Subject
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev    Corr
## (Intercept) 27.127675 (Intr)
## Days         5.903701 0.025
## Residual    47.661517
##
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | Subject
##  Parameter estimates:
##       308       309       310       330       331       332       333
## 1.0000000 0.1859084 0.2584717 0.4806793 0.4968392 1.2163445 0.2606111
##       334       335       337       349       350       351       352
## 0.4261464 0.2417913 0.3401410 0.2957765 0.5228126 0.4717177 0.5260808
##       369       370       371       372
## 0.3288478 0.5252883 0.5155399 0.2355753
## Fixed effects: Reaction ~ Days
##                 Value Std.Error  DF  t-value p-value
## (Intercept) 251.94620  7.054487 161 35.71432       0
## Days         10.26396  1.500955 161  6.83828       0
##  Correlation:
##      (Intr)
## Days -0.105
##
## Standardized Within-Group Residuals:
##          Min           Q1          Med           Q3          Max
```

```
## -2.206546213 -0.562641914  0.002539438  0.673593045  2.301717727
##
## Number of Observations: 180
## Number of Groups: 18

# compound symmetry
sleepModLmeCompSymR<-lme(Reaction~Days,
                 random=~Days|Subject,
                 correlation=corCompSymm(form=~1|Subject),
                 data=sleepstudy)
summary(sleepModLmeCompSymR)
## Linear mixed-effects model fit by REML
##  Data: sleepstudy
##        AIC      BIC    logLik
##   1757.628 1779.901 -871.8141
##
## Random effects:
##  Formula: ~Days | Subject
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev    Corr
## (Intercept) 24.740241 (Intr)
## Days         5.922103 0.066
## Residual    25.591843
##
## Correlation Structure: Compound symmetry
##  Formula: ~1 | Subject
##  Parameter estimate(s):
##           Rho
## -1.249001e-17
## Fixed effects: Reaction ~ Days
##                 Value Std.Error  DF  t-value p-value
## (Intercept) 251.40510  6.824516 161 36.83853       0
## Days         10.46729  1.545783 161  6.77151       0
##  Correlation:
##      (Intr)
## Days -0.138
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -3.95355735 -0.46339976  0.02311783  0.46339621  5.17925089
##
## Number of Observations: 180
## Number of Groups: 18
# virtually no difference as the correlation parameter for the compound symmetry is essentially 0
```

Further reading

For some more examples on how to specify different structures for both random effects and residuals, see, e.g.

- West, B.T., Welch, K.B., Galecki, A.T. (2015) *Linear Mixed Models. A Practical Guide Using Statistical Software.*, 2nd ed., CRC Press
- https://rpsychologist.com/r-guide-longitudinal-lme-lmer

Exercise 2

Download the dataset `autism.csv` from GitHub.

This dataset was collected by University of Michigan researchers and has data from a prospective cohort study of 214 children. The file you downloaded is a subset of 158 children with autism spectrum disorder.

The dependent variable is VSAE - Vineland Socialisation Age Equivalent - a combined, numerical score that includes assessment of interpersonal relationships, play/leisure time activities and coping skills.

Language development was assessed using the Sequenced Inventory of Communication Development scale and children were classified according to this (variable `sicdegp`).

The other two variables in the dataset are the child's age (`age`) at each visit and the child ID (`childid`).

Explore the dataset and develop a model for `vsae`.

Exercise 2 (Solution)

Let's read in the data and check what the data looks like:

```
autism<-read.csv("autism.csv")
dim(autism) # 612 observations, 4 variables
## [1] 612   4
head(autism) # variables are age, csae, sicdegp, childid
##   age vsae sicdegp childid
## 1   2    6       3       1
## 2   3    7       3       1
## 3   5   18       3       1
## 4   9   25       3       1
## 5  13   27       3       1
## 6   2   17       3       3
```

Let's look at the distributions of the various variables:

Age:

```
table(autism$age)
##
##   2   3   5   9  13
## 156 150  91 120  95
```

SICD scores:

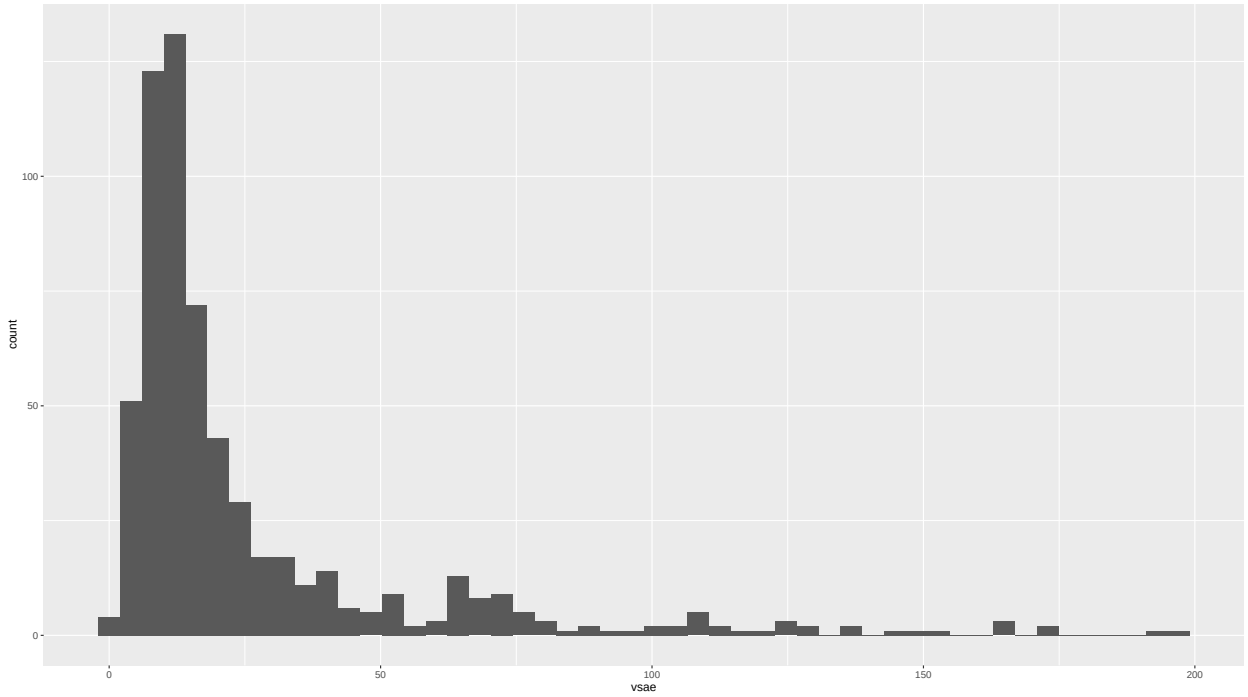
```
table(autism$sicdegp)
##
##   1   2   3
## 192 255 165
```

Number of observations for each child:

```
table(table(autism$childid))
##
##  1  2  3  4  5
##  2 14 28 72 42
```

VSAE:

```
ggplot(data=autism,mapping=aes(x=vsae)) +
  geom_histogram(bins=50)
```

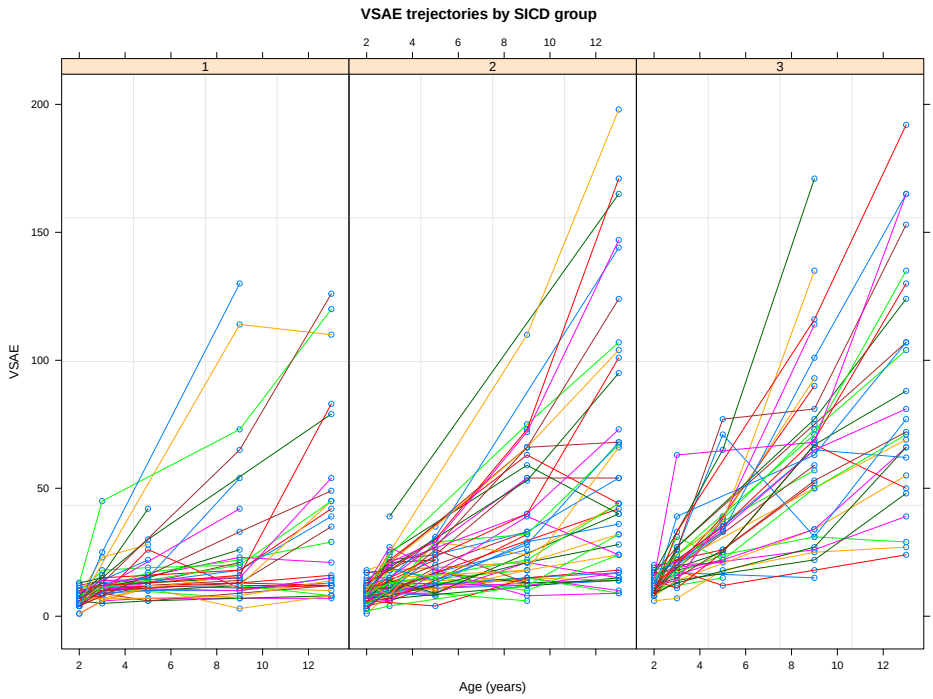


There seem to be observations with missing VSAE. We will remove those (there are only 2 missing values, so should not be an issue; if a substantial number were missing, we would need to be more careful about what we do about the missing values).

```
autism<-autism[!is.na(autism$vsae),]
dim(autism)
## [1] 610   4
```

Let's plot trajectories for `vsae` by SICD group.

```
autism.gr<-groupedData(vsae ~ age | childid, outer = ~ as.factor(sicdegp), data=autism) # function from
plot(autism.gr, display="childid", outer=T, aspect=2, key=F,
     xlab="Age (years)", ylab="VSAE", main="VSAE trejectories by SICD group")
```



From this visual inspection it seems that:

- VSAE increases with age.
- Some differences between SICD groups (e.g. in group 1, it seems like there are more 'flat' trajectories).

You might also argue that the relationship between `age` and `vsae` gets steeper as `vsae` increases. This could suggest a parabolic curve (i.e. involving a squared term). This is not very pronounced however and so, for simplicity, we do not consider this here for this exercise.

From this we will start with a model for `vsae` that includes fixed effects for `age`, `sicdegp`, interaction between `age` and SICD group, and two random effects associated with childid: random intercept, random age effect.

First a bit more reformatting:

```
autism$sicdegp<-factor(autism$sicdegp) # to avoid this being treated as a numeric score
```

Now we can fit a first model.

```
mod1<-lmer(vsae ~ age + sicdegp + age:sicdegp + # fixed
                  (age | childid),               # random
           REML=T,
           data=autism)
## boundary (singular) fit: see ?isSingular

summary(mod1)
## Linear mixed model fit by REML ['lmerMod']
## Formula: vsae ~ age + sicdegp + age:sicdegp + (age | childid)
##    Data: autism
##
## REML criterion at convergence: 4695.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.8687 -0.3519 -0.0231  0.3134  5.2901
```

```
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  childid  (Intercept) 74.14    8.610
##           age         16.21    4.026    -1.00
##  Residual             60.44    7.774
## Number of obs: 610, groups:  childid, 158
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)    1.8396     1.6913   1.088
## age            2.9728     0.6277   4.736
## sicdegp2      -0.3219     2.2336  -0.144
## sicdegp3      -3.8533     2.4794  -1.554
## age:sicdegp2   0.7142     0.8301   0.860
## age:sicdegp3   4.3334     0.9194   4.713
##
## Correlation of Fixed Effects:
##             (Intr) age    scdgp2 scdgp3 ag:sc2
## age         -0.893
## sicdegp2    -0.757  0.676
## sicdegp3    -0.682  0.609  0.517
## age:sicdgp2  0.675 -0.756 -0.893 -0.460
## age:sicdgp3  0.609 -0.683 -0.461 -0.889  0.516
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

This model is singular, suggesting too many model parameters. Particularly the correlation between the 2 random effects (intercept and slope) is -1, suggesting too many random effects. Since there are clearly very different slopes for different participants and that the overall level does not seem to vary too much between individual, we can drop the random intercept and only include a random slope.

```
mod2<-lmer(vsae ~ age + sicdegp + age:sicdegp + # fixed
                (0 + age | childid),           # random; the `0` means no intercept
          REML=T,
          data=autism)

summary(mod2)
## Linear mixed model fit by REML ['lmerMod']
## Formula: vsae ~ age + sicdegp + age:sicdegp + (0 + age | childid)
##    Data: autism
##
## REML criterion at convergence: 4854.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.5107 -0.3949  0.0214  0.3788  4.5033
##
## Random effects:
##  Groups   Name Variance Std.Dev.
##  childid  age   8.198   2.863
##  Residual      84.532   9.194
## Number of obs: 610, groups:  childid, 158
##
```

```
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)    2.4816     1.2711   1.952
## age            2.8219     0.4698   6.006
## sicdegp2      -1.2934     1.6732  -0.773
## sicdegp3      -4.2301     1.8621  -2.272
## age:sicdegp2   0.9845     0.6194   1.589
## age:sicdegp3   4.4628     0.6885   6.482
##
## Correlation of Fixed Effects:
##             (Intr) age    scdgp2 scdgp3 ag:sc2
## age         -0.406
## sicdegp2    -0.760  0.309
## sicdegp3    -0.683  0.277  0.519
## age:sicdgp2  0.308 -0.759 -0.403 -0.210
## age:sicdgp3  0.277 -0.682 -0.211 -0.386  0.518
```

The model now converges.

Let's see if we should drop the fixed effects for the interaction terms between age and sicdegp.

```
mod3<-lmer(vsae ~ age + sicdegp +      # fixed
               (0 + age | childid),  # random; the `0` means no intercept
          REML=T,
          data=autism)

summary(mod3)
## Linear mixed model fit by REML ['lmerMod']
## Formula: vsae ~ age + sicdegp + (0 + age | childid)
##    Data: autism
##
## REML criterion at convergence: 4897.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.4623 -0.4209  0.0144  0.4003  4.4037
##
## Random effects:
##  Groups   Name Variance Std.Dev.
##  childid  age  10.44    3.231
##  Residual      85.59    9.252
## Number of obs: 610, groups:  childid, 158
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   0.9736     1.2194   0.798
## age           4.4309     0.2887  15.346
## sicdegp2     -0.3187     1.5680  -0.203
## sicdegp3     -0.3427     1.7556  -0.195
##
## Correlation of Fixed Effects:
##          (Intr) age    scdgp2
## age      -0.221
## sicdegp2 -0.741  0.004
## sicdegp3 -0.664  0.016  0.514
```

We can formally do a likelihood ratio test. Note that for this we need to use maximum likelihood estimation, so we first need to refit the models:

```
mod2ml<-lmer(vsae~age+sicdegp+age:sicdegp+(0+age|childid),REML=F,data=autism)
mod3ml<-lmer(vsae~age+sicdegp+(0+age|childid),REML=F,data=autism)

anova(mod3ml,mod2ml)
## Data: autism
## Models:
## mod3ml: vsae ~ age + sicdegp + (0 + age | childid)
## mod2ml: vsae ~ age + sicdegp + age:sicdegp + (0 + age | childid)
##        Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## mod3ml  6 4915.7 4942.1 -2451.8   4903.7
## mod2ml  8 4877.2 4912.5 -2430.6   4861.2 42.425      2   6.13e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model with the interaction terms has lower AIC than the model without these terms and the p-value from the likelihood ratio test is also extremely low, and hence we keep the interaction terms in the model.

For the purpose of this exercise, our final model is `mod2`:

$$\text{vsae}_{ti} = \beta_0 + \beta_1 \cdot \text{age}_{ti} + \beta_3 \cdot \text{SICD\_2}_i + \beta_4 \cdot \text{SICD\_3}_i + \beta_5 \cdot \text{age}_{ti} \cdot \text{SICD\_2}_i + \beta_6 \cdot \text{age}_{ti} \cdot \text{SICD\_3}_i + u_i \cdot \text{age} + \epsilon_{ti}$$

where t indexes the observation time point and i the individual.

For a more thorough discussion (including adding an age-squared term), please refer to Chapter 6 of West, B.T., Welch, K.B., Galecki, A.T. (2015) *Linear Mixed Models. A Practical Guide Using Statistical Software.*, 2nd ed., CRC Press.