Linear Mixed-Effects Models: Application to Autism Data

Gabriel Afriyie

13/04/2020

Abstract

Linear mixed-effects (LME) modeling is a widely used statistical method for analyzing repeated measures or longitudinal data. In this project, autism was chosen as a topic of research since it is a very important health that needs attention. Vineland Socialization Age Equivalent (vsae) and its predictors are evaluated using LME models. To properly visualize the dataset, graphics are produced in both R and Tableau softwares. The LME model with interaction terms is fit for the autism data. For this analysis, we decided to remove the child-specific random effects associated with the intercept because of the estimation problems, and because there was little variability in the initial VSAE scores for these autistic children at 0 years of age, The final model shows significance of the quadratic time trend and interaction. Model diagnostics revealed that the data required some transformation to ensure normality of residuals and random effects.

Section 1: Introduction

Longitudinal data require that subjects in the study be repeatedly measured across time (Diggle, Heagerty, Liang, & Zeger, 2002; Hedeker & Gibbons, 2006; Vonesh & Chinchilli, 1997). This is the crucial difference between longitudinal data and cross-sectional data, which measures only a single outcome for each individual (Diggle et al., 2002). The outcome measured in longitudinal data may be continuous, binary, ordinal, or categorical in nature.

The two most commonly used approaches to analyzing longitudinal data are referred to as marginal models (population-averaged) and random-effects (subject-specific) models. The random-effects model, on the other hand, considers that regression coefficients vary across individuals (Diggle et al., 2002); a process that stems from the assumption that repeated observations are correlated. In basic terms, there is an average regression coefficient from which each individual deviates given person-specific conditions. Longitudinal data, a special case of repeated measures data, arecharacterized as having both between-subject and within-subject variation, timedependent covariates and missing data (Davis, 2002). Linear mixed-effects model can accommodate these complex features of longitudinal data whereas traditional methods are limited by statistical assumptions. More importantly, the approach allows for explicit modeling of the variation between subjects and within subjects.

This project illustrates fitting linear mixed effects models to the autism data. The data comes from researchers at the University of Michigan as part of a prospective longitudinal study of 214 children. We have a data frame of 612 observations on the following 4 variables:

- *age*: Age in years (2, 3, 5, 9, 13); the time variable.
- *vsae*: Vineland Socialization Age Equivalent: parent-reported socialization, the dependent variable measured at each age.
- *sicdegp*: Sequenced Inventory of Communication Development Expressive Group: categorized expressive language score at age 2 years (1 = Low, 2 = Medium, 3 = High).

- *childid*: Unique child identifier.

In the next section, we present data manipulation and graphically explore the autism data. Section 3 describes how the linear mixed model for the autism data was fit. Section 4 discusses the results and findings. We perform model diagnostics in Section 5 and add conclusions and recommendations in Section 6.

Section 2: Data Manipulation

We use the *dplyr()* package in R software to subtract 2 from age so that 2 years becomes age=0, as reference age, and hence intercept for the age trend in the linear model makes sense. Note that we also change the numerical variable *sicdegp* to a categorical variable with 3 levels. The levels of *sicdegp* are then renamed as "Low", "Medium" and "High".

We produce graphical description of the data in both Tableau and R. We generate Figure 1 by plotting *vsae* against *age* and split the graph by *childid* and *sicdegp*. We color by *childid*. The plots of the observed VSAE values for individual children in Figure 1 show substantial variation from child to child within each level of SICD group; the VSAE scores of some children tend to increase as the children get older, whereas the scores for other children remain relatively constant. On the other hand, we do not see much variability in the initial values of VSAE at age 0 years for any of the levels of SICD group. Therefore, we attribute this variation entirely to random error rather than to between-subject variability and we will fit our by removing the random child-specific intercepts, while retaining the same fixed effects and the child-specific random effects of age and age-squared.

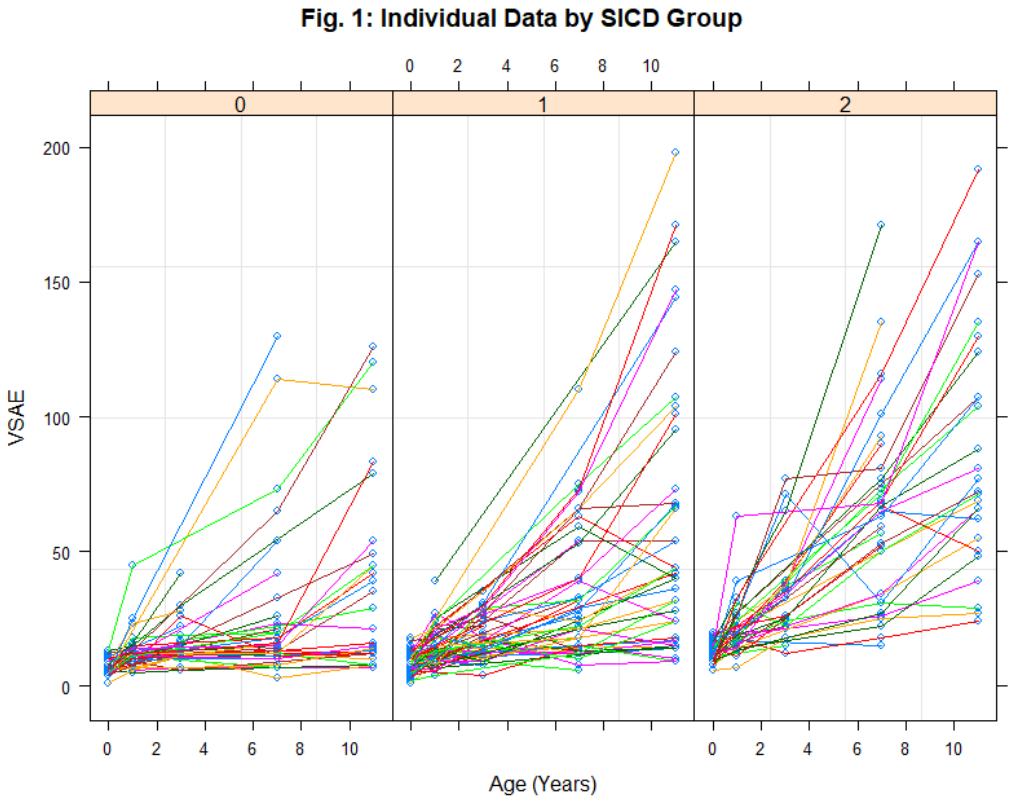


Figure 1: VSAE values plotted against age for children in each SICD group.

Section 3: Model Specification and Analysis

With VSAE as the response variable, we fit a quadratic regression model for each child, which includes the fixed effects of age, age-squared, SICD group, the SICD group by age interaction, and the SICD group by age-squared interaction. As already mentioned in the previous section, We also include two random effects associated with each child: a random age effect, and a random age-squared effect, without a random intercept. We fit the model of this form:

$$vsae = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 sicdegp + \beta_4 age * sicdegp + + \beta_5 age^2 * sicdegp + random(age + age^2 - 1|childid) + \varepsilon$$

The first portion of the model formula, $vsae = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 sicdegp + \beta_4 age * sicdegp + + \beta_5 age^2 * sicdegp$, defines the continuous response variable (VSAE), and the terms that have fixed effects in the model (including the interactions). The second portion of the model formula, $+random(age + age^2 - 1|childid)$, indicates the variables that have random effects associated with them (in parentheses). The random effects and residuals are all considered to be normally distribted with mean zero and respective variances. We use the *lmer()* function in R and SPSS's mixed procedure to fit the full model. We then perform model selection by passing the results through the *step()* function of the "lmertest" R package.

Results and Findings

After fitting the full model stated in the previous section, the *step()* function is used to select the model with the least AIC. This selected model is labeled 'fit2' in the Appendix. In this model, the SICD group by age-squared interaction is dropped. We summarize the selected model as follows:

The variance explained by the random effects are different from 0. This means that the random effects matter and a regular linear model would be inappropriate. The mean predicted VSAE score for children at 0 years of age in the reference category of the SICD group (SICDEGP = Low) is 8.47591. The effect of one unit increase in a child's age is to increase their *vsae* by 2.08 units, assuming any other combination of factor levels. The quadratic trend associated with time (which represented in the variable "age2", shows a significant increase in *vsae* by 0.109 units.

Moreover, for children in the High *sicdegp* group, *vsae* increases over the followup by 4.99 units as compared to the other *sicdegp* groups. The *sicdegp* groups, Low and Medium are not statistically different. The interaction term is also significant. We can also find the correlation of fixed effects in the output. The fixed effects and negatively and/or uncorrelated. We test the significance of the random effect variances by the *ranova()* R function. The output shows that both variances are highly significant. We now do some model diagnosis in the next section.

Diagnostics for the Final Model

We now check the assumptions for our selected model fitted using REML estimation, using informal graphical procedures in the R software. We first assess the assumption of normality for the residuals in our model. This is done using the *qqnorm()* function to create normal plots of residuals.

Figure 3 does not show a straight line. This means that the normality assumption is violated. We perform log transformation on the response variable (vsae) to correct this. Figure 4 shows a straight line. This means that the residuals are in the model are now normally distributed. Similarly, we check the assumption of normality for the random effects.

Figure 5 also shows that the random effects are normally distributed even without transforming the data.

Conclusion

Linear mixed-effects modeling is a powerful approach to modeling longitudinal data, such as the autism data. This approach has the ability to model both between-subject and within-subject variability through random-effects. Both time-invariant and time-variant covariates can be accommodated in the model. Our main findings suggest that the *vsae* score in the high *sicdegp* increases over the followup by 4.99 units as compared to the other *sicdegp* groups.

Model selection is rarely a perfect and bias-free process. With linear mixed-effects modeling, researchers have to be concerned not only with selecting appropriate covariates for the model, but also choosing the best covariance structure for the random-effects. Given the limitations in the modeling approach and the study design, weare confident in the final model presented. We feel that our findings will significantly contribute to the study and research of autism in children.

Reference

Oti, R., Anderson, D., Risi, S., Pickles, A. & Lord, C., Social Trajectories Among Individuals with Autism Spectrum Disorders, Developmental Psychopathology (under review), 2006.

West, B., Welch, K. & Galecki, A, Linear Mixed Models: A Practical Guide Using Statistical Software, Chapman Hall / CRC Press, first edition, 2006.

Linear Mixed-Effects Models: Applications to the Behavioral Sciences and Adolescent CommunityHealth Lizmarie Gabriela Maldonado University of South Florida, lmaldon3@alumni.health.usf.edu

Davis, C. S. (2002). Statistical Methods for the Analysis of Repeated Measurements. New York, NY: Springer.

Diggle, P., Heagerty, P., Liang, K.Y., & Zeger, S. (2002). Analysis of Longitudinal Data. New York, NY: Oxford University Press.

Hedeker, D., & Gibbons, R.D. (2006). Longitudinal Data Analysis. Hoboken, NJ: John Wiley & Sons, Inc.

Vonesh, E. F., & Chinchilli, V. M. (1997). Linear and nonlinear models for the analysis of repeated measurements. New York, NY: Marcel Dekker, Inc.

Appendix

(Analysis via R:)

We load the libraries that was required for the analysis.

```
library(stringr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(WWGbook)
library(klaR)
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

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

```
library(lme4)
```

```
## Loading required package: Matrix
```

```
library(lmerTest)
```

```
##
## Attaching package: 'lmerTest'
```

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

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

```
##     step
```

```
library(lattice)
library(nlme)
```

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

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

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

Attaching the dataset to be used.

```
attach(autism)
```

Recoding the variables as described in Section 2.

```
df = autism
df = df %>% mutate(age=age - 2, sicdegp = sicdegp - 1)

df = df %>% mutate(sicdegp=as.factor(sicdegp))

df = df %>% mutate(sicdegp=recode(sicdegp,
                                  "2"="High",
                                  "1"="Medium",
                                  "0"="Low"))
```

We saved the new dataset by writing it to a csv file.

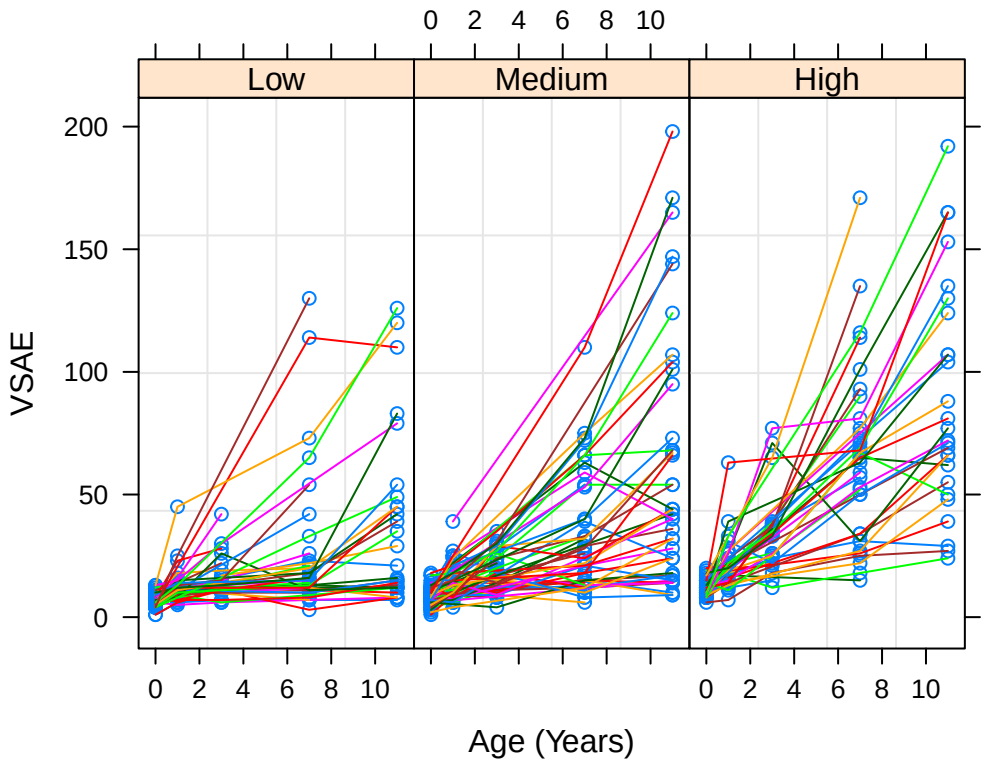
```
write.csv(df, file = "C:/Users/gafri/Desktop/Winter 2020/Statistical Data Analysis W20/Final project/au
```

Creating figure 1 (Plotting vsae score for each child in within the 3 sicdegp group)

```
autism.gl <- groupedData(vsae ~ age | childid, outer = ~ sicdegp, data = df)
 # Generate individual profiles in Figure 1.
   plot(autism.gl, display = "childid", outer = TRUE, aspect = 2,
         key = F, xlab = "Age (Years)", ylab = "VSAE",

            main = "Fig. 1: Individual Data by SICD Group")
```

Fig. 1: Individual Data by SICD Group



Fitting the linear mixed model

```
age2=df$age*df$age

fit = lmer(vsae ~ age + age2 + sicdegp + age*sicdegp + age2*sicdegp + (age + age2 -1 | childid), df)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00214699 (tol = 0.002, component 1)
```

```
summary(fit)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: vsae ~ age + age2 + sicdegp + age * sicdegp + age2 * sicdegp +
##     (age + age2 - 1 | childid)
##    Data: df
##
## REML criterion at convergence: 4615.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.2235 -0.3794 -0.0501  0.2891  6.8864
##
## Random effects:
##  Groups   Name Variance Std.Dev. Corr
##  childid  age  14.6668  3.8297
##           age2  0.1315  0.3626   -0.32
##  Residual      38.4991  6.2048
```

```
## Number of obs: 610, groups:  childid, 158
##
## Fixed effects:
##                      Estimate Std. Error        df t value Pr(>|t|)
## (Intercept)        8.354e+00  7.353e-01 3.480e+02  11.361  < 2e-16 ***
## age                2.307e+00  7.493e-01 1.946e+02   3.079  0.00238 **
## age2               6.938e-02  7.870e-02 1.271e+02   0.882  0.37967
## sicdegpMedium      1.378e+00  9.722e-01 3.481e+02   1.418  0.15713
## sicdegpHigh        5.416e+00  1.094e+00 3.444e+02   4.953 1.15e-06 ***
## age:sicdegpMedium  5.499e-01  9.930e-01 1.927e+02   0.554  0.58038
## age:sicdegpHigh    3.296e+00  1.092e+00 1.856e+02   3.020  0.00288 **
## age2:sicdegpMedium 4.917e-03  1.033e-01 1.256e+02   0.048  0.96210
## age2:sicdegpHigh   1.346e-01  1.134e-01 1.234e+02   1.187  0.23740
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) age    age2   scdgpM scdgpH ag:scM ag:scH ag2:sM
## age         -0.435
## age2         0.326 -0.598
## sicdegpMedm -0.756  0.329 -0.246
## sicdegpHigh -0.672  0.292 -0.219  0.509
## ag:scdgpMdm  0.328 -0.755  0.452 -0.435 -0.221
## ag:scdgpHgh  0.299 -0.686  0.411 -0.226 -0.426  0.518
## ag2:scdgpMd -0.248  0.456 -0.762  0.325  0.167 -0.598 -0.313
## ag2:scdgpHg -0.226  0.415 -0.694  0.171  0.321 -0.313 -0.592  0.529
## convergence code: 0
## Model failed to converge with max|grad| = 0.00214699 (tol = 0.002, component 1)
```

Stepwise model selection

```
s1 = step(fit)

s1
```

```
## Backward reduced random-effect table:
##
##                                    Eliminated npar  logLik    AIC     LRT Df
## <none>                                             13 -2307.6 4641.3
## age in (age + age2 - 1 | childid)           0   11 -2364.8 4751.7 114.404  2
## age2 in (age + age2 - 1 | childid)          0   11 -2349.6 4721.2  83.927  2
##                                    Pr(>Chisq)
## <none>
## age in (age + age2 - 1 | childid)   < 2.2e-16 ***
## age2 in (age + age2 - 1 | childid)  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Backward reduced fixed-effect table:
## Degrees of freedom method: Satterthwaite
##
##              Eliminated Sum Sq Mean Sq NumDF  DenDF F value    Pr(>F)
## age2:sicdegp          1  72.27   36.14     2 123.52  0.9386   0.39393
## age2                  0 251.80  251.80     1 124.36  6.4923   0.01205 *
## age:sicdegp           0 974.31  487.15     2 174.55 12.5605 8.008e-06 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Model found:
## vsae ~ age + age2 + sicdegp + (age + age2 - 1 | childid) + age:sicdegp
```

```
fit2=lmer(vsae ~ age + age2 + sicdegp + (age + age2 - 1 | childid) + age:sicdegp,df)

summary(fit2)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: vsae ~ age + age2 + sicdegp + (age + age2 - 1 | childid) + age:sicdegp
##    Data: df
##
## REML criterion at convergence: 4611.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.2656 -0.3976 -0.0545  0.2914  6.8472
##
## Random effects:
##  Groups   Name Variance Std.Dev. Corr
##  childid  age  14.5316  3.8120
##           age2  0.1265  0.3557   -0.31
##  Residual      38.7843  6.2277
## Number of obs: 610, groups:  childid, 158
##
## Fixed effects:
##                     Estimate Std. Error        df t value Pr(>|t|)
## (Intercept)          8.47591    0.70940 383.33580  11.948  < 2e-16 ***
## age                  2.08070    0.64830 222.27160   3.209  0.00153 **
## age2                 0.10901    0.04278 124.36483   2.548  0.01205 *
## sicdegpMedium        1.36480    0.92153 400.48222   1.481  0.13939
## sicdegpHigh          4.98766    1.03784 397.98381   4.806 2.19e-06 ***
## age:sicdegpMedium    0.57252    0.79610 177.67744   0.719  0.47299
## age:sicdegpHigh      4.06812    0.87986 172.66417   4.624 7.36e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) age    age2   scdgpM scdgpH ag:scM
## age         -0.359
## age2         0.188 -0.378
## sicdegpMedm -0.742  0.221  0.002
## sicdegpHigh -0.659  0.195  0.005  0.508
## ag:scdgpMdm  0.234 -0.696 -0.006 -0.318 -0.161
## ag:scdgpHgh  0.213 -0.632  0.000 -0.164 -0.309  0.514
```

Testing the significance of random effects of variances of selected model.

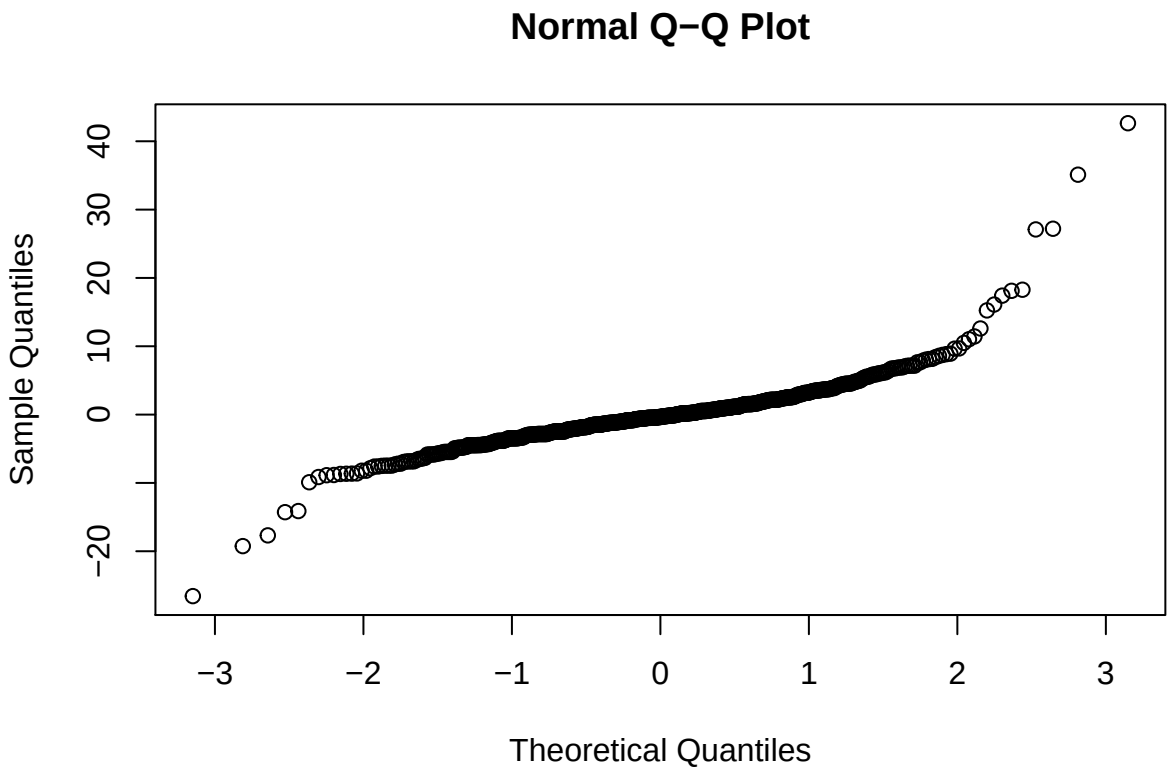
```
ranova(fit2)
```

```
## ANOVA-like table for random-effects: Single term deletions
##
## Model:
```

```
## vsae ~ age + age2 + sicdegp + (age + age2 - 1 | childid) + age:sicdegp
##                                    npar  logLik    AIC     LRT Df Pr(>Chisq)
## <none>                               11 -2305.8 4633.6
## age in (age + age2 - 1 | childid)     9 -2363.2 4744.3 114.751  2  < 2.2e-16
## age2 in (age + age2 - 1 | childid)    9 -2346.3 4710.7  81.103  2  < 2.2e-16
##
## <none>
## age in (age + age2 - 1 | childid)  ***
## age2 in (age + age2 - 1 | childid) ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Diagnostics

```
qqnorm(residuals(fit2))
```



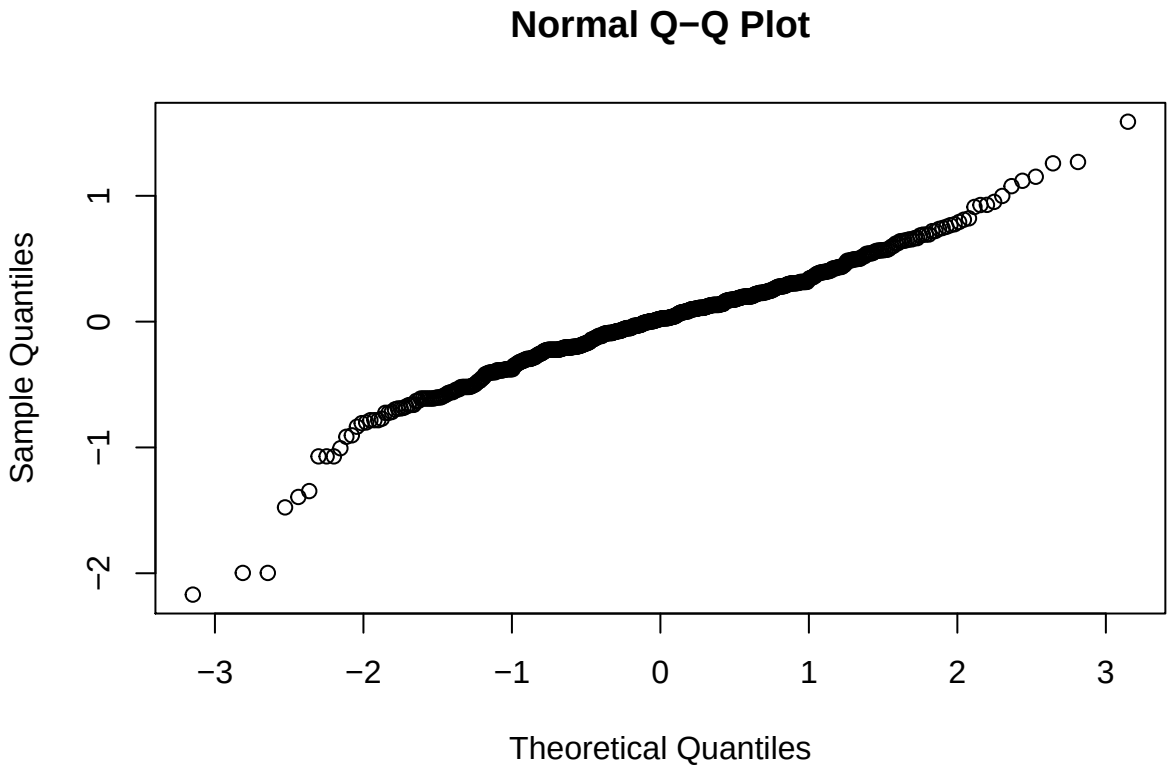

```
logvsae = log(df$vsae)

fit3=lmer(logvsae ~ age + age2 + sicdegp + (age + age2 - 1 | childid) + age:sicdegp,df)
```

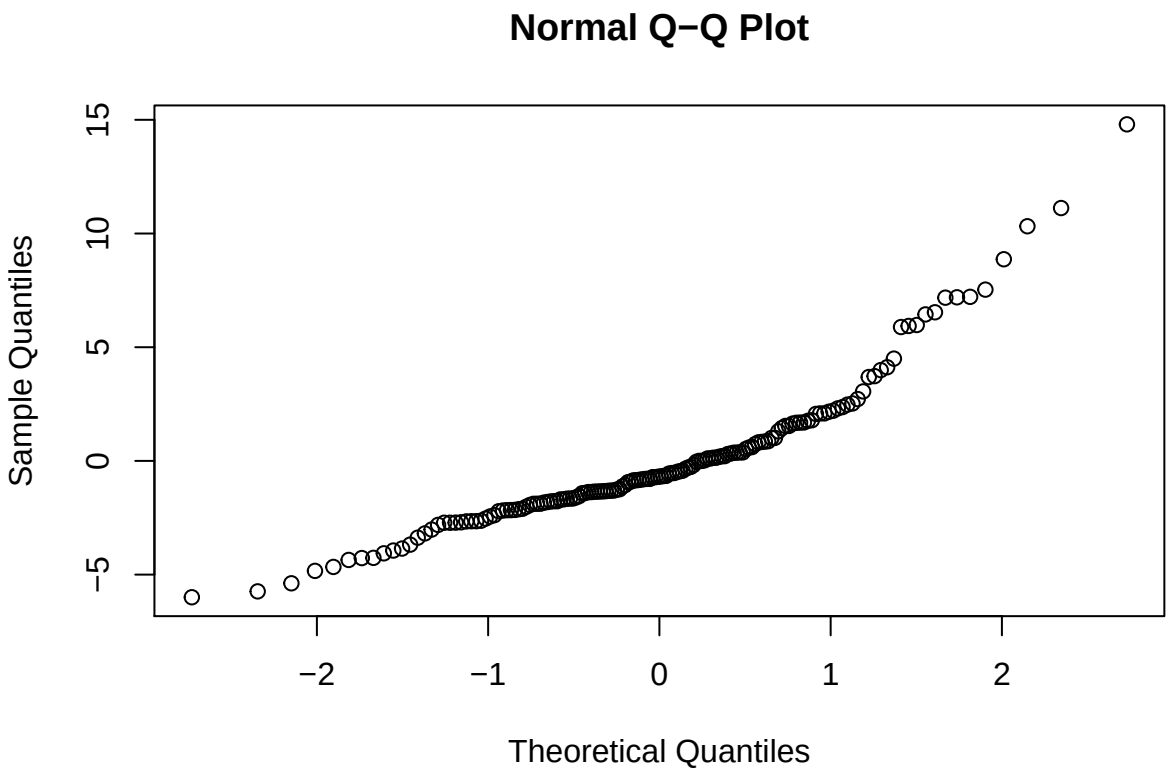
```
## boundary (singular) fit: see ?isSingular
```

```
## Warning: Model failed to converge with 1 negative eigenvalue: -9.3e+03
```

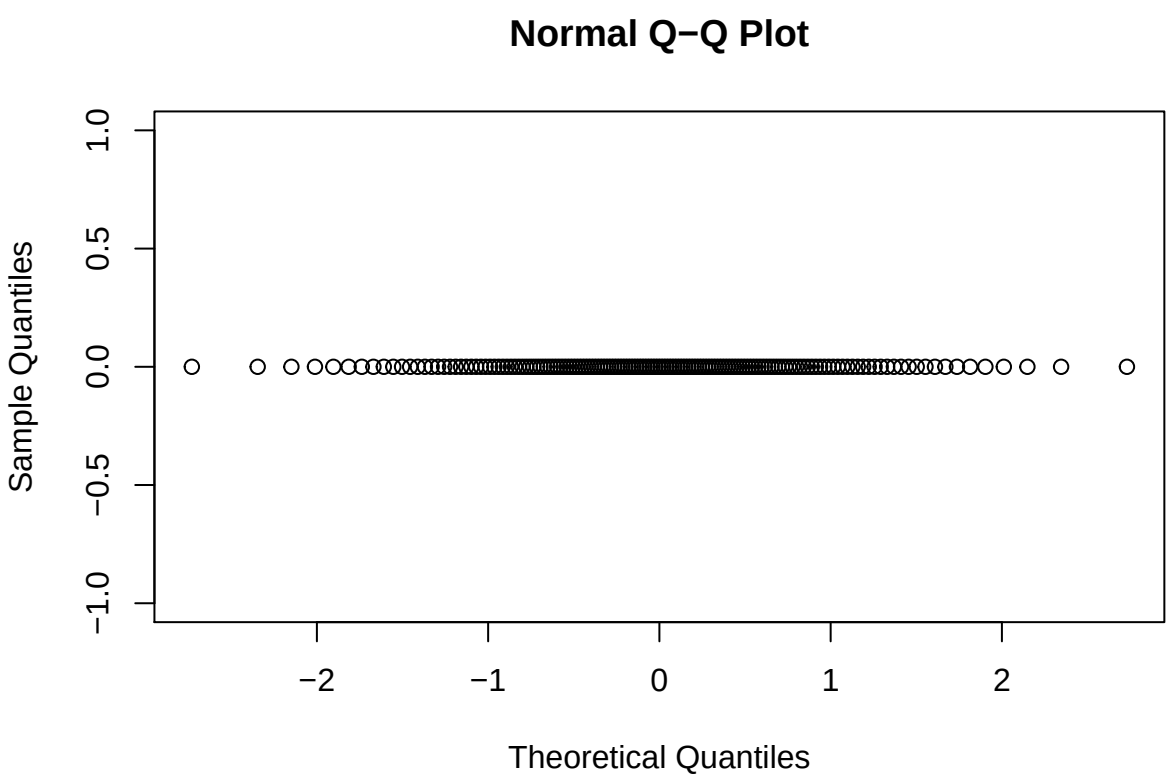
```r
qqnorm(residuals(fit3))
```




```r
qqnorm(ranef(fit2)$childid[[1]])
```




```
qqnorm(ranef(fit3)$childid[[1]])
```


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
1.0
0.5
0.0
−0.5
−1.0
−2
−1
0
1
2