MLM Assignment 2

Christine Hedde-von Westernhagen Emilia Löscher Kyuri Park

Utrecht University, 09 March, 2022

Data Description

In this assignment, we analyze the `curran_wide.csv` data, which contains the information about the age, antisocial behavior, reading skills, emotional support, cognitive stimulation, and mother's age of 221 sampled children. Antisocial behavior and reading skills are measured over 4 occasions. In this analysis, we do not use the variables for child's gender and emotional support.

The (pre-processed) data specifics are as follows:

- `id`: child id
- `time`: measurement occasion ranging from 0 to 3
- `anti`: antisocial behavior (time-variant)
- `read`: reading recognition skills (time-variant & grand-mean centered)
- `momage`: mother's age measured at the first occasion (time-invariant & grand-mean centered)
- `homecog`: cognitive stimulation measured at the first occasion (time-invariant & grand-mean centered)

1. Convert the wide data file into a long format. Check the data and recode if necessary.

As seen below, the original data (`curran_wide.csv`) is converted into a long format.
`time` is recoded from 1 - 4 to 0 - 3 such that the first measurement is set to 0.
`read`, `momage`, and `homecog` are centered at the grand mean such that the the intercept represents the expected value of antisocial behavior for an average child (with average reading skills + average mom's age + average cognitive stimulation) when these predictors are included in the model.
Maybe we should center each when adding them.. not sure about centering read if it makes any sense

```
## # A tibble: 6 x 13
##      id anti1 anti2 anti3 anti4 read1 read2 read3 read4   sex momage homecog
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>   <dbl>
## 1    34     3     6     4     5   2.1   2.9   4.5   4.5     1     28       9
## 2    58     0     2     0     1   2.3   4.5   4.2   4.6     0     28       9
## 3   125     1     1     2     1   2.3   3.8   4.3   6.2     0     29      10
## 4   133     3     4     3     5   1.8   2.6   4.1   4       1     28       8
## 5   163     5     4     5     5   3.5   4.8   5.8   7.5     1     28      10
## 6   248     1     2     2     0   3.5   5.7   7     6.9     0     28       9
## # ... with 1 more variable: homeemo <dbl>
```

```
## # A tibble: 6 x 6
##      id  time  anti   read momage homecog
##   <dbl> <dbl> <dbl>  <dbl>  <dbl>   <dbl>
## 1    34     0     3 -2.25    2.40 -0.0995
## 2    34     1     6 -1.45    2.40 -0.0995
## 3    34     2     4  0.155   2.40 -0.0995
## 4    34     3     5  0.155   2.40 -0.0995
## 5    58     0     0 -2.05    2.40 -0.0995
## 6    58     1     2  0.155   2.40 -0.0995
```

Table 1

Descriptive statistics

	n	mean	sd	median	min	max	skew	se
id	884	3679	2495	3410	34	8870	0.39	83.92
time	884	1.5	1.12	1.5	0	3	0	0.04
anti	884	1.82	1.82	1	0	10	1.12	0.06
read	884	0	1.62	0.05	-3.65	4.05	0.11	0.05
momage	884	0	1.87	0.4	-4.6	3.4	-0.14	0.06
homecog	884	0	2.45	-0.1	-6.1	4.9	-0.37	0.08

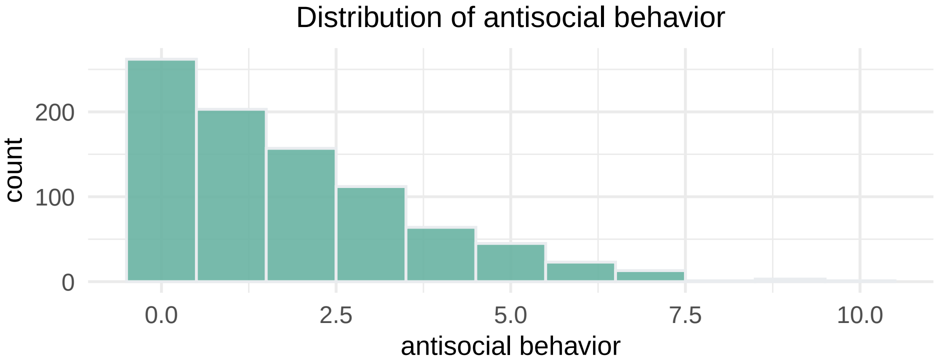


Figure 1: Right-skewed Antisocial behavior

Table 1 shows the descriptive statistics of each variable in the re-coded data. There are total 884 obervations. The mean values of `read`, `momage`, and `homecog` are 0, as expected, and they seem to be not overly skewed given that they have small skewness (i.e., <|1|). The skewness value of `anti` is higher than 1, which indicates that the distribution is skewed. Hence, we checked on the distribution of `anti` and as shown in *Figure 1*, it is right-skewed where most values are clustered around the left tail. This is also aligned with the (top-left) boxplot shown in *Figure 2*.

1a. Check the linearity assumption, report and include plots.

Based on *Figure 3*, it is concluded that the linearity assumption is met in both level 1 and level 2, even though the relationships between the variables seem to be rather weak (i.e., slopes are quite flat), except for the one *(d) Cognitive stimulation - Average Antisocial behavior.*

1b. Check for outliers.

Figure 2 shows that *antisocial behavior* has several univariate outliers and *cognitive stimulation* also has one outlier in the left tail. In *Figure 3*, we can check the bivariate outliers at each level. At level 1, although there are few points spotted in the upper-side of plots *(a)* and *(b)*, they do not seem to be influential. Also, at level 2, no such influential outlier is observed.

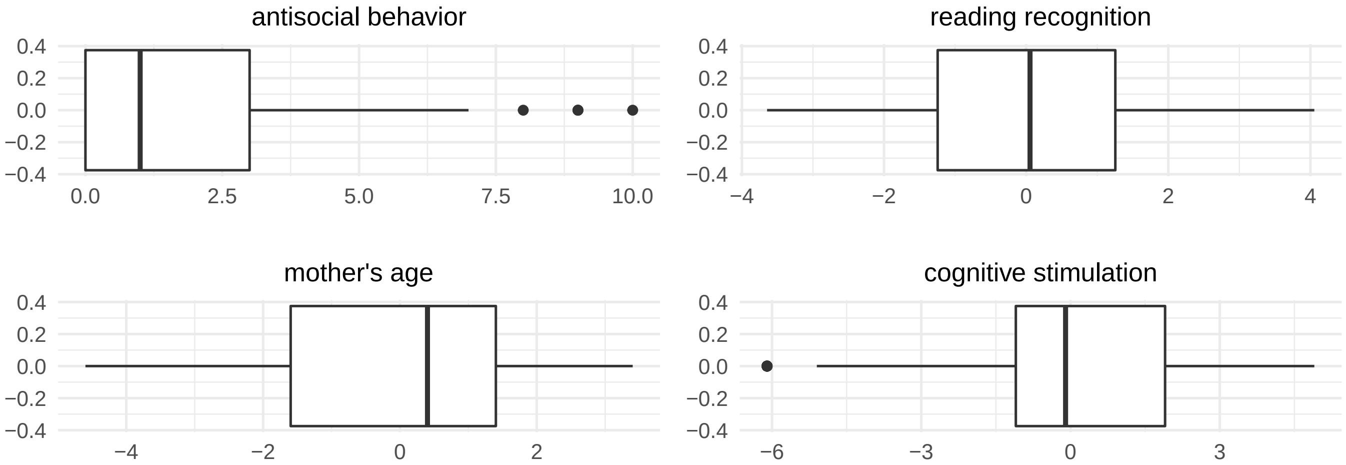


Figure 2: Overall distribution of each variables

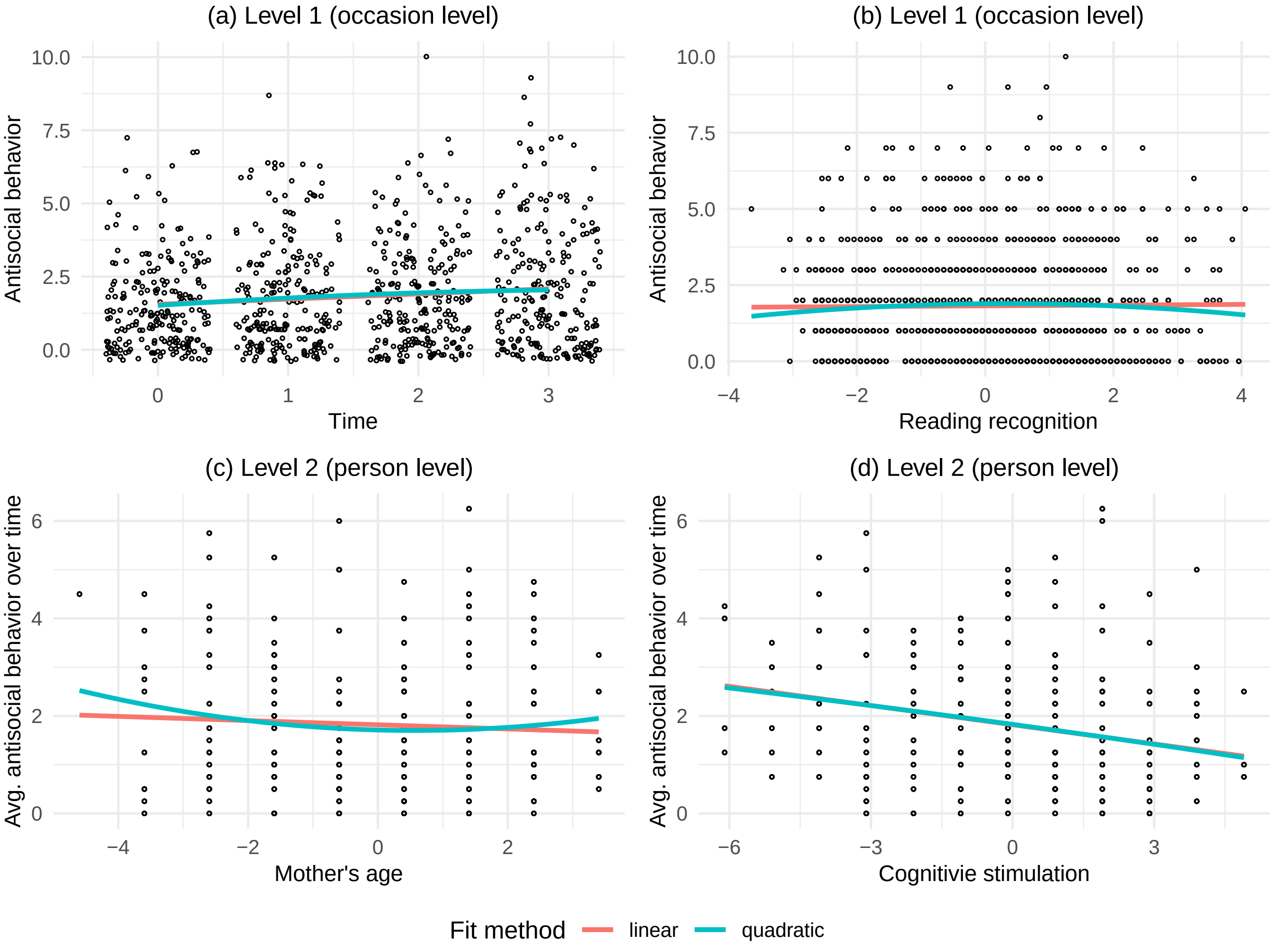


Figure 3: Scatterplots to insepct linearity and outliers

2. Answer the question: should you perform a multilevel analysis?

2a. What is the mixed model equation?

– Mixed Model Equation

$$y_{ti} = \beta_{00} + u_{0i} + e_{ti}$$

- y_{ti} refers to antisocial behavior of child i at time t.
- β_{00} refers to the overall intercept, which is the average antisocial behavior over all children.
- u_{0i} refers to the random residual error at the person level (level 2), which represents the deviation from the overall intercept (β_{00}) of child i.
- e_{ti} refers to the residual error at the occasion level (level 1).

2b. Provide and interpret the relevant results.

As shown in *Table 2*, the intercept term is identical as 1.82 in M_0 and M_1. This overall intercept represents the average antisocial behavior across all children. M_1 decomposes the variance term in a variance at level 1 (Var_{occ}) and level 2 (Var_{sub}).

The deviance for the random intercept model turns out to be significantly smaller than that of the single-level model, $\chi^2(1) = 231.97$, $p < .001$, indicating that Var_{occ} is significantly greater than 0. AIC value is also lower with the random intercept model ($AIC_{M_1} = 3343.5$) compared to the single-level model ($AIC_{M_0} = 3573.5$), which is in accordance with the deviance difference test result.

```r
# model0: single-level regression model for comparison
model0 <- lm(anti ~ 1, data = curran_long)
summary(model0)

# model1: random intercept model ((benchmark model to compute ICC))
model1 <- lmer(anti ~ 1 + (1|id), REML = FALSE, data = curran_long)
summary(model1)

# check the significance of random intercept
anova(model1, model0)
```

Table 2

Single-level model and Intercept-only model

Model	M_0: single-level model	M_1: random intercept
Fixed part	Coefficient(SE)	Coefficient(SE)
Intercept	1.82(.06)	1.82(.10)
Random part		
Var_{occ}	3.32	1.74
Var_{sub}		1.58
Deviance	3569.5	3357.5
AIC	3573.5	3343.5
Deviance difference[a]		231.97***

Note

$^{***}p < .001, ^{**}p < .01$

[a] p-value for χ^2 test is one-sided p-value.

2c. What is the intraclass correlation?

The intraclass correlation (ρ) is calculated as follows:

$$\rho = \frac{\sigma^2_{u0}}{\sigma^2_{u0} + \sigma^2_e}$$

```
ICC <- 1.579/(1.579+1.741)
```

The ICC in this model is equal to $\rho = 0.48$.

2d. What is your conclusion regarding the overall question regarding the necessity of performing a multilevel analysis?

Yes, we should perform the multilevel analysis in this case, because not only the data structure is nested (i.e., multiple measurements within each individual), but also the difference between individuals accounts for about *48%* of the total variance. In other words, the intraclass correlation – ICC: the proportion of the total variance explained by the between-individual differences – is *0.476*, which is high enough that the multilevel analysis is warranted.

In addition, *Figure 4* shows the relationships between *antisocial behavior - time* and *antisocial behavior - reading skills* in each children (first 20). Here we see that the slopes and intercepts vary across the children, which again suggests the necessity of multilevel analysis that can properly address this between-person variabilities.

3. Add the time-varying predictor(s).

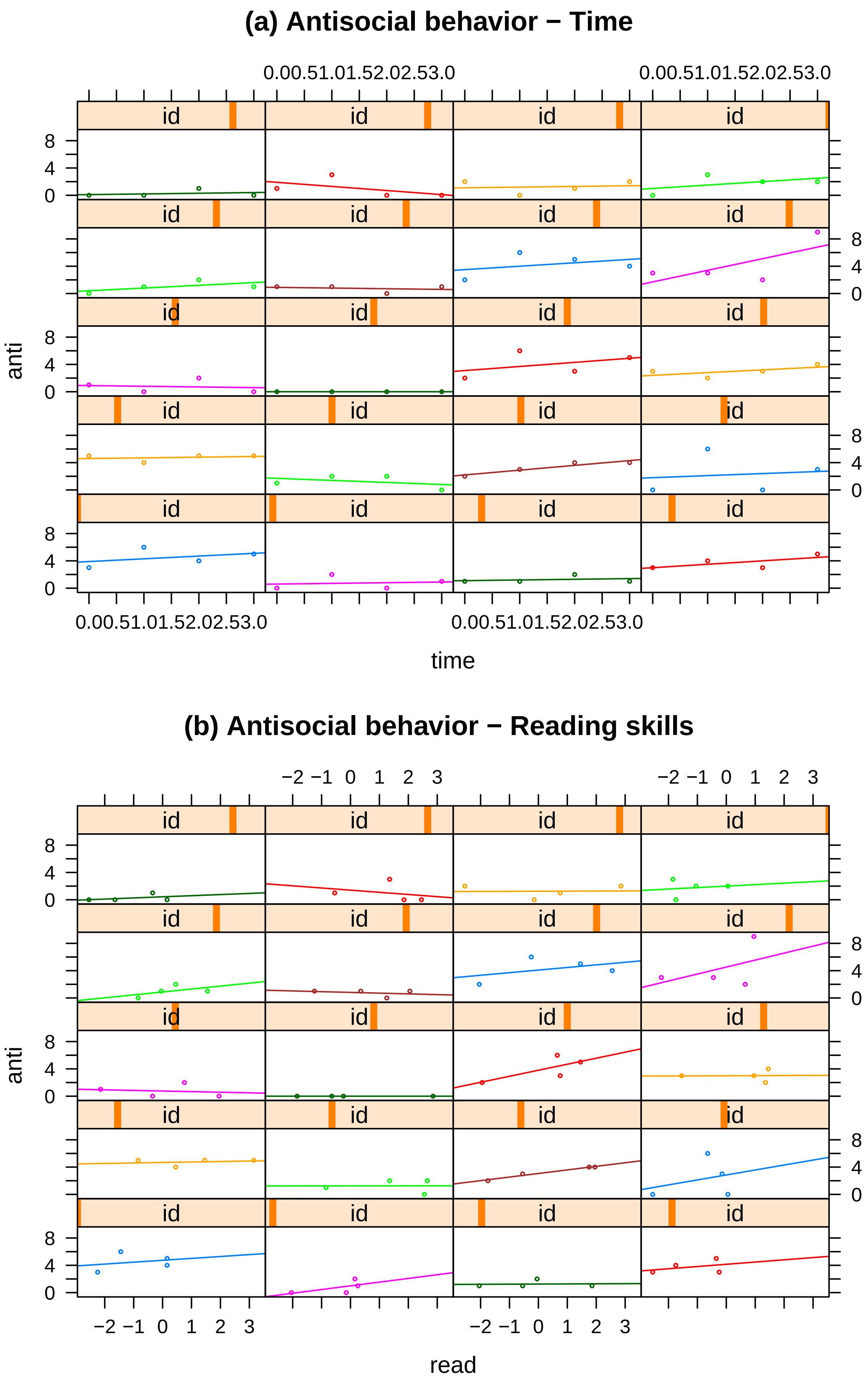
Provide and interpret the relevant results and provide your overall conclusion.

- mention centering technique <- I forgot what this comment was about exactly . . . .

As shown in *Table 3*, the intercept term is now 1.55 in M_2 with `time` predictor, which refers to the average antisocial behavior at the first occasion (time = 0). `time` turns out to be a statistically significant predictor of `anti`, $b_{time} = 0.18, t(663) = 4.513, p < .001$, meaning that antisocial behavior is expected to increase by 0.18 on each succeeding occasion. By adding `time`, the occasion-level residual error variance (Var_{occ}) goes down to 1.69, but the subject-level variance (Var_{sub}) goes up to 1.60. This is because in the longitudinal data with fixed occasion measurements, there is no variation in time-points between subjects. Subsequently, Var_{occ} is overestimated and Var_{sub} is underestimated in the intercept-only model, M_1. In order to correct this, measurement occasion variable needs to be included in the model. Hence, M_2 with the `time` predictor correctly decomposes the variance terms and correspondingly can be used as a benchmark model for computing the R^2.

The significance of adding `time` to the random intercept model(M_1):
The deviance of M_2 is significantly smaller than the deviance of M_1, $\chi^2(1) = 20.06$, $p < .001$ (see *Table 2* and *Table 3*). It indicates that M_2 fits significantly better than the intercept-only model, M_1. AIC also corresponds to this, as $AIC_{M_2} = 3325.5$ is lower than $AIC_{M_1} = 3343.5$.

In M_3 with `time` and `read` together, the `time` is still a significant predictor $b_{time} = 0.21, t(882) = 2.73, p < .01$, but `read` is not a significant predictor of `anti`, $b_{read} = -0.03, t(831) = -0.54, p = .588$. The deviance of M_3 is 3317.2, which is almost the same as M_2, and the difference in the deviances is not significant as expected, $\chi^2(1) = 0.29$, $p = .297$. Also, AIC of M_3 is slightly higher than AIC of M_2 (see *Table 3*), which again suggests that adding `read` does not provide a better fit. Therefore, we drop `read` and proceed with M_2 including only the `time` predictor.



(a) Antisocial behavior − Time

(b) Antisocial behavior − Reading skills

Figure 4: Scatterplots of children to inspect the variabilities in intercept/slopes

```r
# model2: add a time-varying predictor, time ((benchmark model for computing R2))
model2 <- lmer(anti ~ 1 + time + (1|id), REML = FALSE, data = curran_long)
summary(model2)
anova(model2, model1)

# model3: add a time-varying predictor, read
model3 <- lmer(anti ~ 1 + time + read + (1|id), REML = FALSE, data = curran_long)
summary(model3)
anova(model3, model2)
```

Table 3

Adding time-varying predictors

Model	M_2: add time	M_3: add read
Fixed part	Coefficient(SE)	Coefficient(SE)
Intercept	1.55(.11)	1.50(.15)
time	0.18(.04)***	0.21(.07)**
read		-0.03(.06)
Random part		
Var_{occ}	1.69	1.69
Var_{sub}	1.60	1.58
Deviance	3317.5	3317.2
AIC	3325.5	3327.2
Deviance difference[a]	20.06***	0.29

Note

$^{***}p < .001,^{**}p < .01$

[a] p-value for χ^2 test is one-sided p-value.

4. On which level or levels can you expect explained variance?

Calculate and interpret the explained variances.

In theory, we can expect the explained variances (R^2) at both occasion (level 1) and subject level (level 2), as the level 1 predictor in general can explain the variances in both levels. However, in this case where the additional time-varying predictor `read` is not significant and correspondingly does not reduce the variance at the occasion level (but rather increase it slightly), $R^2_{occasion}$ becomes negative and therefore cannot be defined.

```r
m2var.lv1 <- 1.689 # level 1 variance in model 2 (benchmark model)
m2var.lv2 <- 1.592 # level 2 variance in model 2 (benchmark model)
m3var.lv1 <- 1.693 # level 1 variance in model 3
m3var.lv2 <- 1.576 # level 2 variance in model 3

# explained variance at level 1 (occasion level)
R2.lv1 <- (m2var.lv1 - m3var.lv1) / m2var.lv1
R2.lv1

# explained variance at level 2 (subject level)
R2.lv2 <- (m2var.lv2 - m3var.lv2) / m2var.lv2
R2.lv2
```

The computed R^2 values for each level using M_2 as our benchmark:

- $R^2_{occasion} =$ not defined
- $R^2_{subject} = 0.0101$

It means that reading recognition skill (`read`) further explains about 1% of the variance between children.

5. Add the time invariant predictor(s) to the model.

Provide and interpret the relevant results and provide your overall conclusion.

In M_{4a} we add both time invariant predictors, mother's age (`momage`) and cognitive stimulation (`homecog`). As shown in *Table 4*, `momage` is turned out to be not significant, $b_{momage} = -0.001, t(221) = -0.02, p = .985$, and only `homecog` is significant, $b_{homecog} = -0.13, t(221) = -3.35, p < .001$. This means that for each one point increase in `homecog` (the more cognitive stimulation a child receives), the average `anti` is expected to decrease by 0.13 (the average antisocial behavior of a child goes down). The intercept term value remains almost the same as the previous model (1.55), but the interpretation now with the added predictors is the average antisocial behavior for a child with the average cognitive stimulation and average mom's age at the first occasion (as `homecog` and `momage` are centered at the grand mean).
By adding these level 2 predictors, we see that the variance at level 2 (subject level) is decreased and this correspondingly produces $R^2_{subject} = .065$.
The significance of adding `momage` and `homecog` to M_2:
The deviance of M_{4a} is significantly smaller than the deviance of M_2, $\chi^2(2) = 11.642$, $p < .01$. It indicates that M_{4a} fits significantly better than M_2 with only `time`. The lower $AIC_{M_{4a}} = 3317.8$ also suggests that M_{4a} is preferable.

However, given that `momage` is not a significant predictor, we drop `momage` and only include `homecog` in the model, M_{4b}. The fixed part of model does not change much, both `time`, $b_{time} = 0.18, t(663) = 4.513, p < .001$, and `homecog` are significant,$b_{homecog} = -0.13, t(221) = -3.35, p < .001$. The interpretation of `homecog` is the same as described above, but the intercept term would now represent the average antisocial behavior for a child with the average cognitive stimulation at the first occasion.
By keeping only `homecog` in the model, we see that the variance at level 2 (subject level) is decreased by the same amount as in M_{4a} and accordingly results in the same $R^2_{subject} = .065$.
The significance of adding only `homecog` to M_2:
The deviance of M_{4b} is significantly smaller than the deviance of M_2, $\chi^2(1) = 11.641$, $p < .001$. It indicates that M_{4b} fits significantly better than M_2 with only `time`. In addition, $AIC_{M_{4b}} = 3315.8$, which is not only lower than M_2 but also slightly lower than M_{4a}, indicating that M_{4b} is a better fit than M_{4a}. Given these overall results, we decided to proceed with M_{4b} with `time` and `homecog` predictors.

```r
## proceed with a model without the non-significant predictor, 'read'
# model4a: add time-invariant predictors, momage & homecog
model4a <- lmer(anti ~ 1 + time + momage + homecog + (1|id), REML = FALSE, data= curran_long)
summary(model4a)
anova(model4a, model2)

# model4b: remove the non-significant predictor, 'momage'
model4b <- lmer(anti ~ 1 + time + homecog + (1|id), REML = FALSE, data= curran_long)
summary(model4b)
anova(model4b, model2)
```

Table 4

Adding time-invariant predictors

Model	M_{4a}: add momage + homecog	M_{4b}: add only homecog
Fixed part	Coefficient(SE)	Coefficient(SE)
Intercept	1.55(.11)	1.55(.11)
time	0.18(.04)***	0.18(.04)***
momage	0.00(.05)	
homecog	-0.13(.04)***	-0.13(.01)***
Random part		
Var_{occ}	1.70	1.69
Var_{sub}	1.50	1.49
Deviance	3305.8	3305.8
AIC	3317.8	3315.8
Deviance difference^{ab}	11.64**	11.64***

Note

$^{***}p < .001, ^{**} p < .01$

[a] p-value for χ^2 test is one-sided p-value.

[b] Here the deviance is compared with M_2.

6. On which level or levels can you expect explained variance?

Calculate and interpret the explained variances.

We can expect the explained variances (R^2) at the subject level (level 2), as the level 2 predictor can only explain the variance in level 2.

```
m2var.lv1 <- 1.689 # level 1 variance in model 2 (benchmark model)
m2var.lv2 <- 1.592 # level 2 variance in model 2 (benchmark model)
m4var.lv1 <- 1.689 # level 1 variance in model 4b
m4var.lv2 <- 1.488 # level 2 variance in model 4b

# explained variance at level 2 (subject level)
R2.lv2 <- (m2var.lv2 - m4var.lv2) / m2var.lv2
R2.lv2
```

The computed R^2 value for the subject level using M_2 as our benchmark:

- $R^2_{subject} = 0.0653$

It means that cognitive stimulation provided at home (`homecog`) explains about 6% of the variance between children.

7. For the time-varying predictor(s), check if the slope is fixed or random.

7a. What are the null- and alternative hypotheses?

- H_0: $\sigma^2_{u1} = 0$; The slope for the time variable is equal across the children.
- H_1: $\sigma^2_{u1} > 0$; The slope for the time variable varies across the children.

7b. Provide and interpret the relevant results.

- interpret result chi2 test

```
# model5: let 'time' have random slopes
model5 <- lmer(anti ~ 1 + time + homecog + (1 + time|id), REML = FALSE, data = curran_long)
summary(model5)
anova(model5, model4b)

# # model5b: let 'read' have random slopes
# model5b <- lmer(anti ~ 1 + time + homecog + (1 + read|id),
#                 REML = FALSE, data= curran_long)
# summary(model5b)
#
# model5b2 <- lmer(anti ~ 1 + time + homecog + read + (1 + read|id),
#                 REML = FALSE, data= curran_long)
# summary(model5b2)
#
# anova(model5b, model4)
#
# # model5c: let both have random slopes --> model fails to converge
# model5c <-  lmer(anti ~ 1 + time + homecog + (1 + time + read|id),
#                  REML = FALSE, data = curran_long)
```

Table 5

Model with random slope and covariance of intercept and slope

Model	M_5: random slope
Fixed part	Coefficient(SE)
Intercept	1.55 (.10)
time	0.18(.04)***
homecog	-0.10(.04)**
Random part	
Var_{occ}	0.95
Var_{sub}	1.53
Var_{time}	0.10
$Cor_{sub*time}$	0.41
Deviance	3279.3
AIC	3293.3
Deviance difference^{ab}	26.56***

Note

$^{***}p < .001, ^{**}p < .01$

^{a} p-value for χ^2 test is one-sided p-value.

^{b} Here the deviance is compared with M_{4b}.

7c. Provide an overall conclusion.

The variance of time is significant when added to the model therefore we conclude that the effect of…

8. If there is a random slope, set up a model that predicts the slope variation.

We have found that there is a random slope for the variable `time`. In the following, we are fitting two models to check whether the variable `momage` or the variable `homecog` can predict the slope variation.

Provide and interpret the relevant results and provide your overall conclusion.

From model 6a, it can be seen that homecog can predict the slope variation of time as the interaction of homecog and time is significant ($p = 0.008$). Note, however, that the fixed effect for homecog is not significant anymore after including the interaction term ($p = 0.1056$). Nevertheless, the fixed effect of homecog is not removed from the models as fixed effect that are included part of an interaction should always remain in the model. The estimate for the interaction term is -0.04532. This means that for children who have more cognitive stimulation (higher values for homecog), their antisocial behavior increases less strongly over time. This is visualized in the figure below.

In model 6b the interaction of time and momage was added to the model. From the summary of model 6b, it can be seen that momage cannot be used to predict the slope variation of time as the interaction of momage and time is not significant ($p = 0.88749$). Furthermore, the fixed effect of momage is also not significant ($p = 0.90528$). This had already been the case in model 4a and, therefore, momage had been removed from the model. It was added here once more because it is part of the interaction.

```r
## model6: add a cross-level interaction
# check if homecog can explain the time variance
model6a <- lmer(anti ~ 1 + time + homecog + homecog*time + (1+time|id),
                REML = FALSE, data = curran_long)
summary(model6a)
anova(model6a, model5)

# check if momage can explain the time variance
model6b <- lmer(anti ~ 1 + time + momage + homecog + momage*time + (1+time|id),
                REML = FALSE, data = curran_long)
summary(model6b)
anova(model6b, model5)
```

Table 6

Adding a cross-level interaction

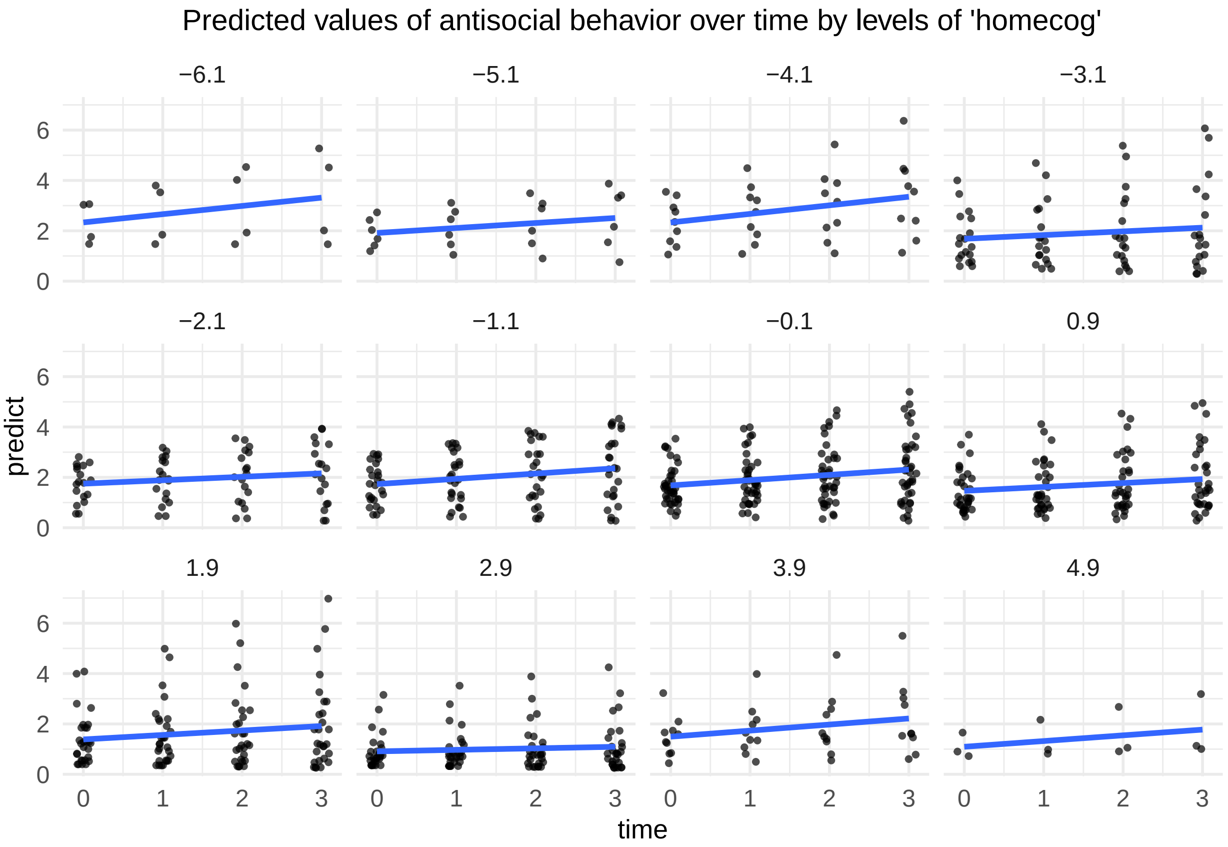
Model	M_{6a}: ‘homecog*time‘	M_{6b}: ‘momage*time‘
Fixed part	Coefficient(SE)	Coefficient(SE)
Intercept	1.55(.10)	1.55(.10)
time	0.18(.04)***	0.18(.04)***
momage		-0.01(.05)
homecog	-0.06(.04)	-0.10(.04)**
time:homecog	-0.05(.02)**	
time:momage		0.00(.02)
Random part		
Var_{occ}	1.53	1.53
Var_{sub}	0.94	0.95
Var_{time}	0.08	
$Cor_{sub*time}$	0.47	0.41
Deviance	3272.4	3279.2
AIC	3288.4	3297.2
Deviance difference^{ab}	6.88**	0.05

Note

$^{***}p < .001,^{**}p < .01$

^{a} p-value for χ^2 test is one-sided p-value.

^{b} Here the deviance is compared with M_5.



-> for children with more homecog, antisocial behavior increases less strongly/decreases over time

So `model6a` is our final model.

9. Decide on a final model.

We choose model 6a to be our final model as it has the best model fit among all the tested models. The level 1 and 2 model equations, as well as the mixed model equation are given below.

9a. Provide the separate level 1 and 2 model equations, as well as the mixed model equation.

– Level 1 Model Equation

$$anti_{ti} = \pi_{0i} + \pi_{1i} time_{ti} + e_{ti}$$

– Level 2 Model Equations

$$\pi_{0i} = \beta_{00} + \beta_{01} homecog_i + u_{0i}$$

$$\pi_{1i} = \beta_{10} + \beta_{11} homecog_i + u_{1i}$$

– Mixed Model Equation

$$anti_{ti} = \beta_{00} + \beta_{10} time_{ti} + \beta_{01} homecog_i + \beta_{11} homecog_i \times time_{ti} + u_{0i} + u_{1i} time_{ti} + e_{ti}$$

9b. Check the normality assumption for both the level-1 and level-2 errors, report.

```r
# level 1 residuals
resid_lvl1 <- residuals(model6a)

# level 2 residuals
resid_lvl2 <- ranef(model6a)$id
```

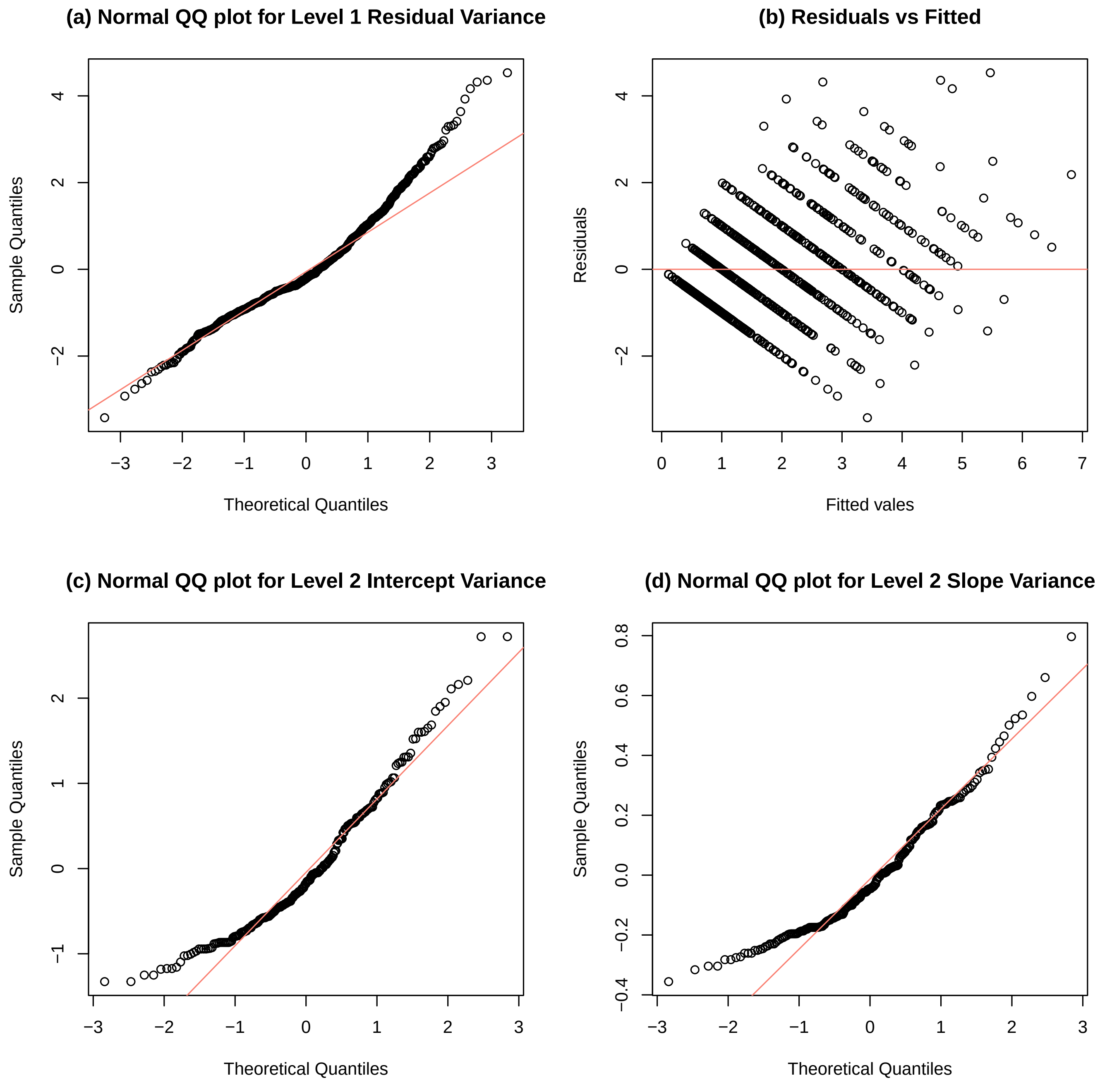


Figure 5: Q-Q plots for level 1 and level 2 residuals

- normality violated, anti very skewed see figure 1, consider log transformation (mention that we checked it)

We checked the normality assumption on level 1 and on level 2. In all three cases, the normality assumption is (somewhat) violated. This can be seen from the respecitve Q-Q plots as the points were not on a straight diagonal line. They were either s-shaped for the level 1 errors or u-shaped for both of the level 2 errors.

The histogram of the outcome variable antisocial behavior (anti) showed that the distribution of the variable is very (right) skewed (see Figure 1). A log-transformation of the variable anti should be considered because this might make the normality assumption more justifiable. When looking at the Q-Q plots for model 6a using the log-transformed outcome variable anti, the points in all three of the Q-Q plots were closer to being on a straight diagonal line than for the non-transformed variable. Hence, it might be a good idea to analyze the data with a log-transformed version of the outcome variable.

Contribution

- Christine: Coding, interpretation, writing
- Emilia: Coding, interpretation, writing, Excel
- Kyuri: Coding, interpretation, writing