MLM-Longitudinal & Contextual(Assignment-2)

Kalyango Jovan

04/03/2022

Libraries

```
library(lme4)
library(foreign)
library(jtools)
library(lmerTest)
library(tidyverse)
```

Data Inspection

The data file curran_wide.csv in R, in each row of the curran_wide.csv file represents a case, and the variables are given in the columns.

```
curran <- read.csv(file = "curran_wide.csv", header=TRUE)
head(curran)
#View(curran)
```

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

The dataset, curran_wide.csv is converted from wide to long format and then the relevant variables are recoded in such a way we include an inter-pretable *zero value*. The dependent variable measuring antisocial behavior **anti** with a meaningful zero value.The variable time, 1 is subtracted from every value in order to have a meaningful zero value(0 to 3).The variables measuring reading skills **read**, cognitive stimulation **homecog** and mother's age **momage** are all centered around their grand mean.

```
curran_longformat <- pivot_longer(data = curran, cols = c('anti1':'read4'),
                          names_to = c(".value", "time"),
                          names_pattern = "(anti|read)(.)")
head(curran_longformat)
curran_long <- curran_longformat %>%
  select(id, momage:homecog, time:read)
```

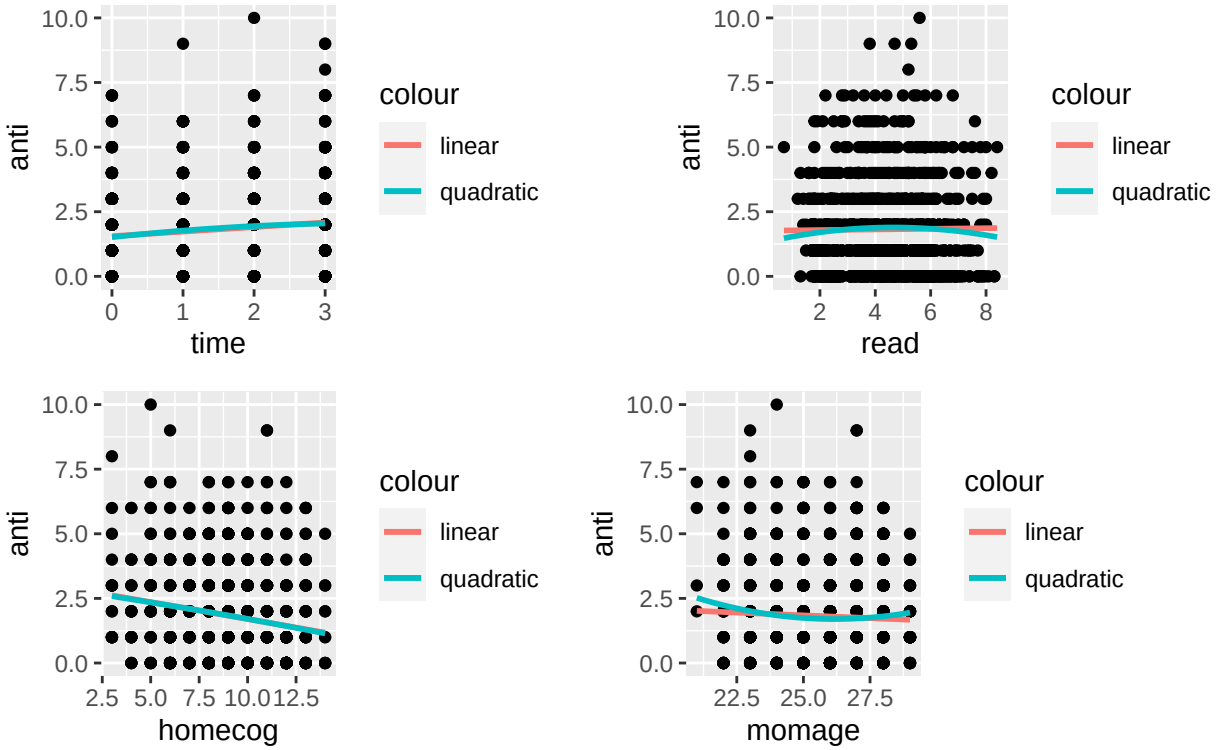
Centering

```
# Centering time so that it includes a meaningful zero point
curran_long$time <- as.numeric(curran_long$time) - 1
# Grand mean centering 'read'
curran_long$readGMC <- curran_long$read - mean(curran_long$read)
```

```
# Grand mean centering 'homecog'
curran_long$homecogGMC <- curran_long$homecog - mean(curran_long$homecog)
# Grand mean centering 'momage'
curran_long$momageGMC <- curran_long$momage - mean(curran_long$momage)
```

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

To check whether there is a linear relationship between the dependent/outcome variable *anti* and the predictors *time, read, homecog, momage*, we look at the scatter plots between the dependent variable against each predictor.



Overall, there is a clear overlap between linear and quadratic lines for *time* and *homecog*, this implies a linear relationship between the two predictors and the dependent variable *anti*.Whereas, *momage* and *read* the overlap is present for some parts, hence less overlap between quadratic and linear. This may be because, there is quadratic relationship with *anti* and the later predictors. However, this possible quadratic seems slightly different from the linear line, thus, the quadratic relationship is not large enough to be included into the model.There is no evidence to support the non-linear relationship between *momage* and *read* and the dependent variable, therefore, we consider to be in linear relationships.

b) Check for outliers (don't perform analyses, just look in the scatterplots), report.

From the scatterplots, there are some students who scored particularly high on the antisocial behaviors score. However, the reason is unknown whether these high scores are the actual true values of these individuals or whether these high scores are due to measurement error.From, the first plot (X = time, Y = antisocial behaviors), it is visible that one student who scores higher at every measurement occasion.Therefore, it seems that these values are within a reasonable value and therefore conclude that there are no outliers.

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

We perform multilevel analysis, because the data set curran_wide.csv is convertible to long format and it is not required to be balanced and there is varying occasions, thus, the analysis choice is multilevel for longitudinal data. We can mix time variant (occasion level) and time invariant (person level) predictors.

a) What is the mixed model equation?

Mixed

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

Where subscript $t = 0, \ldots, T$ denotes the measurement occasion **level 1**, and subscript $i = 1, \ldots, n$ denotes the individuals **level 2**.

b) Provide and interpret the relevant results (don't just copy the output, report the relevant results in APA style).

To determine whether we should perform a multilevel analysis, comparing the intercept only model with the intercept only model that allows for random intercepts is done, If the model including the random intercepts fits data better, then performing a multilevel analysis to these data is done.

Model_0, the intercept only model, display that the $Deviance(2) = 3569.5 and AIC = 3573.5$.From model_0 to model_1, the goodness-of-fit statistics decrease, with $Deviance(3) = 3337.5 and AIC = 3343.5$. Moving from a intercept only to a random intercept model also shows to be a significant improvement of model fit $\chi^2_1 = 231.97, p < 0.001$. The intercepts thus vary significantly across individuals, model_1 has fits the data better so I will procede with the multilevel.

c) What is the intraclass correlation?

$$ICC = 1.579/(1.579 + 1.741) = 0.48$$

Implies that **48 pecent** of the variance in antisocial behavior is variance between the student, and the remaining variance is the variance within students across time.

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

The comparison, between the regular *"intercept only model"* **model_0** with *"random intercept model"* the can conclusion is that the random intercept model fits the model better so it is necessary for these data to perform a multilevel analysis.

The random intercepts is significantly better model fit compared to the "regular" intercept only model $\chi^2_1 = 231.97, p < 0.001$. This implies that the intercepts vary significantly across the different children or students (i.e., the children have different starting scores on antisocial behavior), and multilevel analysis is useful for analyzing this dataset. Also, an intraclass correlation(ICC) of *48 percent* shows that there is dependence of observations, which may result into underestimation of standard errors if the conventional tools/methods for analyzing the data (e.g.,MANOVA, ANOVA or multiple regression). The underestimation of standard errors can result into an inflation of type I errors. So, the problem with dependent observations can be solved by analyzing the data using multilevel analysis.

3.Add the time-varying predictor(s). Provide and interpret relevant results, and provide your overall conclusion.

The time-varying predictors in the model are time and reading score. First, the baseline model for time in order to have a baseline model to calculate explained variance.

In the **model_2**, time is added as a linear predictor with the same coefficient for all students. At the first measurement occasion, the model predicts a value for antisocial behavior of 1.55, which increases with **0.18** at every succeeding measurement occasion. Time is a significant predictor $t = 4.51, p < 0.001$ in this model. Looking at the goodness-of-fit statistics, the $Deviance(4) = 3317.5 and AIC = 3325.5$, we see that both the deviance and the AIC decreased when we moved from the random intercept model **model_1** to the model including time as predictor **model_2**. This improvement in goodness-of-fit of the model also shows to be a significant improvement $\chi^2_1 = 20.06, p < 0.001$. based on the results the conclusion that adding time as a predictor significantly improves the model, and therefore, time will be included in the succeeding models.

Then, the other time-varying is added as predictor after when a baseline model for calculating the explained variance, namely the reading score.

The model **model_3**, we add the time-varying predictor reading skills of the children. The effect of reading skills is non-significant $t = -0.54, p = 0.59$. The goodness-of-fit statistics of **model_3** also do not show a substantial decrease, the AIC increased compared to **model_2** ($Deviance(5) = 3317.2 and AIC(5) = 3327.2$), the difference of deviance test shows to be non-significant $\chi^2_1 = 0.29, p = 0.59$. Since this predictor has no significant effect on antisocial behavior, we will remove it from the preceding models.However, the model is included in the table(*excel file*).But, because of the insignificance of the model, this model is not going to be used to compare other models with for the difference in deviance.

4.On which level or levels can you expect explained variance? Calculate and interpret the relevant results, and provide your overall conclusion.

Explained variance at level 1 (occasion)

$$(1.689 - 1.693)/1.689 = -0.002$$

Explained variance at level 2 (subject)

$$(1.592 - 1.576)/1.592 = 0.010$$

Model_2 is added as a baseline model for calculating the explained variance. Multilevel analysis assumes a hierarchical sampling model,the variability between subjects in the measurement occasion variable is much higher than the hierarchical sampling model assumes. As a result, the random intercept model overestimates the variance at the occasion level **level 1** and underestimates the variance at the subject level **level 2** (Hox et al., 2017).The model including only the time variable as a predictor **model_2** uses this predictor to model the occasion level variance in the dependent variable, antisocial behavior.Hence, **model_2**, is used as a baseline model for calculating the explained variance, so that the variances estimates at the measurement occasions and the subject level are more realistic.

So,the explained variance on both the measurement occasions levels for the time-varying predictor reading score, explain variance at both levels, because reading score is measured at both the measurement occasion level, in other words it is measured at several points in time, and on the subject level the reading score per individual child. The reading score explains **-0.002** of the variance, indicating that the effect of reading score on antisocial behavior does not vary across the different measurement occasions.

The time-varying predictor reading score explains **0.01** of the variance between the school children, that the implying children differ on their reading scores. We already showed that adding the predictor reading score does not significantly improve the model.Also, the predictor does not explain a lot of variance, it even has a negative explained variance on the occasion level. Therefore, reading score can be removed from from the future models.

Question 5. Add the time invariant predictor(s) to the model. Provide and interpret the relevant results, and provide your overall conclusion.

The time invariant predictors are cognitive stimulation and mother's age. They are time invariant because are only measured at the first measurement occasion, the of invariant predictors remains the same over all the measurement occasions.

From **model_4** the effect of mother's age on antisocial behavior is non-significant $t = -0.02, p = 0.98$. Then, the predictor is removed from the model to get **model_4.1**. However, the effect of cognitive stimulation **homecog** show to have a significant effect on antisocial behavior $t = -3.35, p < .001$. The negative relationship (b = -0.13) implies that children who get more cognitive stimulation at home will have a lower score on antisocial behavior.Furthermore, the goodness-of-fit statistics: $Deviance(5) = 3305.8 and AIC = 3315.8$, which are both lower compared to the previous model **model_2**,because we do not compare with **model_3**.Between,the **model_2** to **model_4.1** is a significant improvement in the goodness-of-fit of the model $\chi^2_1 = 11.64, p < 0.001$. Cognitive behavior does have as significant effect, thus, the variable will remain in the succeeding models.

6. On which level or levels can you expect explained variance? Calculate and interpret the explained variances.

Explained variance level 1 (occassion level)

$$(1.69 - 1.689)/1.69 = 0.0006$$

Explained variance level 2 (subject level)

$$(1.592 - 1.488)/1.592 = 0.064$$

The additional of time-invariant predictors, cause a slightly expected additional explained variance on the occasion level **level 1**, and additional explained variance on the subject **level 2**, cognitive stimulation and mother's age are both time-invariant predictors, meaning that they are only measured at the subject level and not at the occasion level, this is reflected in a larger explained variance *in level 2 than in level 1* This is due to the fact that they are only measured once, thus, they do not vary over time. There is a very little **0.0006** additional variance is therefore explained on the measurement occasion level. The model including cognitive stimulation (mother's age is removed from the model) explains **0.064** or 6.4 percent of the variance between the school children, therefore, addition of the predictor cognitive stimulation explains an additional *0.064 – 0.003= 0.061* of the variance compared to model_2.

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

a) a) What are the null- and alternative hypotheses?

The **null hypothesis** is that the slopes for the reading scores and the time variable **do not vary** across children. The **alternative hypothesis** is that the slopes for the reading scores and the time variable **do vary** across children.

b) Provide and interpret relevant results.

We checked whether the time predictor has varying slopes across the children (i.e. do the rates of change differ across the individuals?).To determine whether adding *random slopes* for the time variable is a significant improvement, we compare the current model **model_5** with **model_4.1**,where cognitive behavior was included as *time-invariant* predictor. Comparing **model_5**, that includes the random slopes for time,

with the **model_4.1**, the conclusion is if we move to the current model is a significant improvement $\chi^2_2 = 26.56, p < 0.001$. Addition of random slopes for time improves the model fit, therefore,thus, we assume that time has varying slopes. In other words, the rates of change in antisocial behavior differ across individuals. Then, we can add random slopes for the predictor reading scores. First, if we add random slopes for the variable reading score, the model does not converge.

The absence of convergence indicates that the data does not fit the model well. This may have resulted from poorly fitted observations.Which is an indication that the addition of random slopes for the reading scores is not efficient. Then, by looking at the difference in deviance between the models,the difference is not significant $\chi^2_3 = 0.238, p = 0.97$.

The addition of random slopes for the reading score does not improve the model fit,therefore, no random slopes for the reading scores are included in the succeeding models, this model is added to the table, because it is discussed here in the text.However, this model will not be used to compare other models using difference of deviance.

c) Provide an overall conclusion.

The performed analyses, concludes that the random slopes for **time** should be included,it is significantly improves the model fit.So, random slopes for time is thus better than a model without, this means that the scores on the antisocial behavior scale over time are allowed to differ across the children. Children may have an increase or a decrease in antisocial behavior over time than others. Adding random slopes for the **reading score** would not improve the model fit,therefore, decided to not include random slopes for the reading scores in the succeeding models.

8. If there is a random slope, set up a model that predicts the slope variation. Provide and interpret relevant results, and provide an overall conclusion.

The random slope effect of time could be predicted by individual-level predictors cognitive stimulation and/or mother's age. So adding these predictors as cross-level interactions to see if this is actually the case.

First, the interaction between time and cognitive stimulation **time:homecog** is added to the model. Looking at the relevant goodness-of-fit statistics $Deviance(8) = 3272.4 and AIC = 3288.4$, the decrease in both AIC and Deviance compared to the previous model **model_5**, means including random slopes for time. Comparing two models,the conclusion is that the model **model_7** significantly improve compared to the previous model $\chi^2_1 = 6.88, p < 0.001$.

Second, the interaction between time and mother's age **time:momage** when added to the model.The goodness-of-fit statistics $Deviance(10) = 3272.1, AIC = 3292.1$, the deviance slightly decreased and the AIC increased compared to the previous model **model_5**. This indicates that additional of the interaction between time and mother's age does not improve the model fit. Then, looking at the difference in deviance test, we can conclude that this model does not significantly improve the model $\chi^2_2 = 0.30, p = 0.86$.

So, the addition of the interaction between time and cognitive stimulation means that the amount of change over time for each child also depends on the amount of cognitive stimulation at home. The explained variance for the cross-level interaction for the slopes of time is **(0.096 − 0.084) / 0.096 = 0.125**, which means that **12.5 percent** of the variance explained in the random slopes.

9. Decide on the final model

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

Level 1

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

Level 2

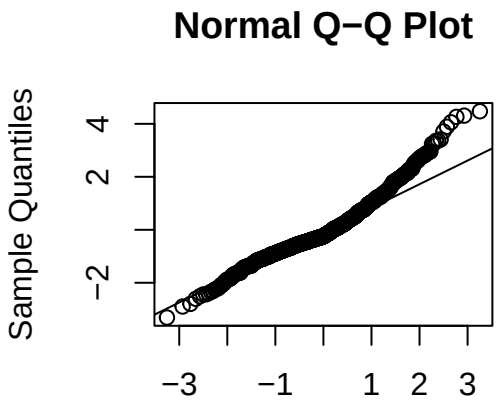
$$\pi_{0i} = \beta_{00} + \beta_{01} * homecog_i + u_{0i}$$
$$\pi_{1i} = \beta_{10} + \beta_{11} * homecog_i + u_{1i}$$

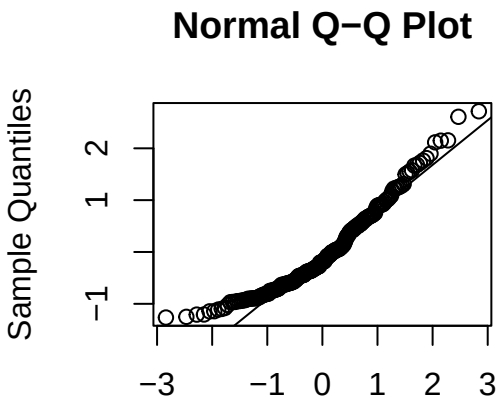
Mixed

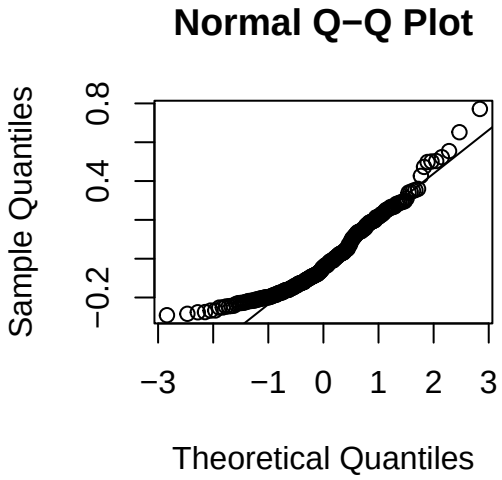
$$anti_{ti} = \beta_{00} + \beta_{10} * time_{ti} + \beta_{01} * homecog_i + \beta_{11} * homecog_i * time_{ti} + u_{1i} + u_{0i} + e_{ti}$$

Where subscript $t = 0, \ldots, T$ denotes the measurement occasion **level 1**, and subscript $i = 1, \ldots, n$ denotes the individuals **level 2**.

b) Check the normality assumption for both the level 1 and level 2 errors, report







For both level 1 and level 2 residual errors there is no evidence to support normality, in other words the assumption is violated. The log transformation, could used to check if there is substantial change to normally

distributed residual errors.From the plots, the starting and the ends points of the plots move away from the diagonal line.

Level 1 error plot,the distribution of the residuals is skewed to the right, since the upper part of the plot is particularly moving away from the diagonal straight line. In the level 2 residuals, the residuals are essentially peaked at the middle of the distribution. The distribution will have a somewhat 'fatter' tail. The second plot for the second level residuals shows that these residuals are skewed to the left, because the lower part of the plot deviates from the straight diagonal line.

The residuals do not lie perfectly on the straight diagonal line of the **Q-Q plots** suggests a possible about the skewness and kurtosis of the distribution. I can finally, conclude that none of these residuals show directly to be perfectly normally distributed.

References

Hox, J. J., Moerbeek, M., & Schoot, V. R. (2017). Multilevel Analysis: Techniques and Applications, Third Edition (Quantitative Methodology Series) (3rd ed.). Routledge.