Longtudinal Analysis

Kalyango Jovan

02/03/2022

Longtudinal Analysis

The multilevel analysis on longitudinal data, is what to practice in this Exercise.

Dataset

The file gpa2.csv contains the data for the GPA example discussed both in the lecture and in the book of Hox, chapter 5.The dependent variable is the GPA score of students, measured during 6 occasions. Besides GPA, we have the number of hours worked in off-campus jobs (Job) on each of these occasions, and the sex (1 = male, 2 = female) and high school GPA of the students.

Libraries used:

```
library(tidyverse)
library(ggplot2)
library(lme4)
library(lmerTest)
```

Data Inspection

The data file GPA2.csv in R, in each row of the gpa2.csv file represents a case, and the variables are given in the columns.

```
GPA <- read.csv(file = "gpa2.csv", header=TRUE)
head(GPA)
```

1. At which level is each variable?

Level 1

Time, gpa, job (both gpa and job are graded/grouped from 1 to 6)

Level 2

student, sex(person specific)

2. Which variable is the level 2 identification variable?

student

3. Convert the wide data file into a file that is suited for multilevel analysis using lme4 in R (that is, we need a long format).

```
library(tidyr)
GPA_long <- pivot_longer(data = GPA, cols = c(4:15),
names_to = c(".value", "time"),
names_pattern = "(gpa|job)(.)")
head(GPA_long)
```

```
## # A tibble: 6 x 6
##   student   sex highgpa time    gpa   job
##     <int> <int>   <dbl> <chr> <dbl> <int>
## 1       1     2     2.8 1       2.3     2
## 2       1     2     2.8 2       2.1     2
## 3       1     2     2.8 3       3       2
## 4       1     2     2.8 4       3       2
## 5       1     2     2.8 5       3       2
## 6       1     2     2.8 6       3.3     2
```

4. Recode Time and Sex.

- Recode the variable Time (from 1-6, to 0-5).
- Recode the variable sex into dummy variables (i.e., using the values 0 and 1).

```
GPA_long$sex <- unclass(GPA_long$sex) - 1
GPA_long$time <- as.numeric(GPA_long$time) - 1
head(GPA_long)
```

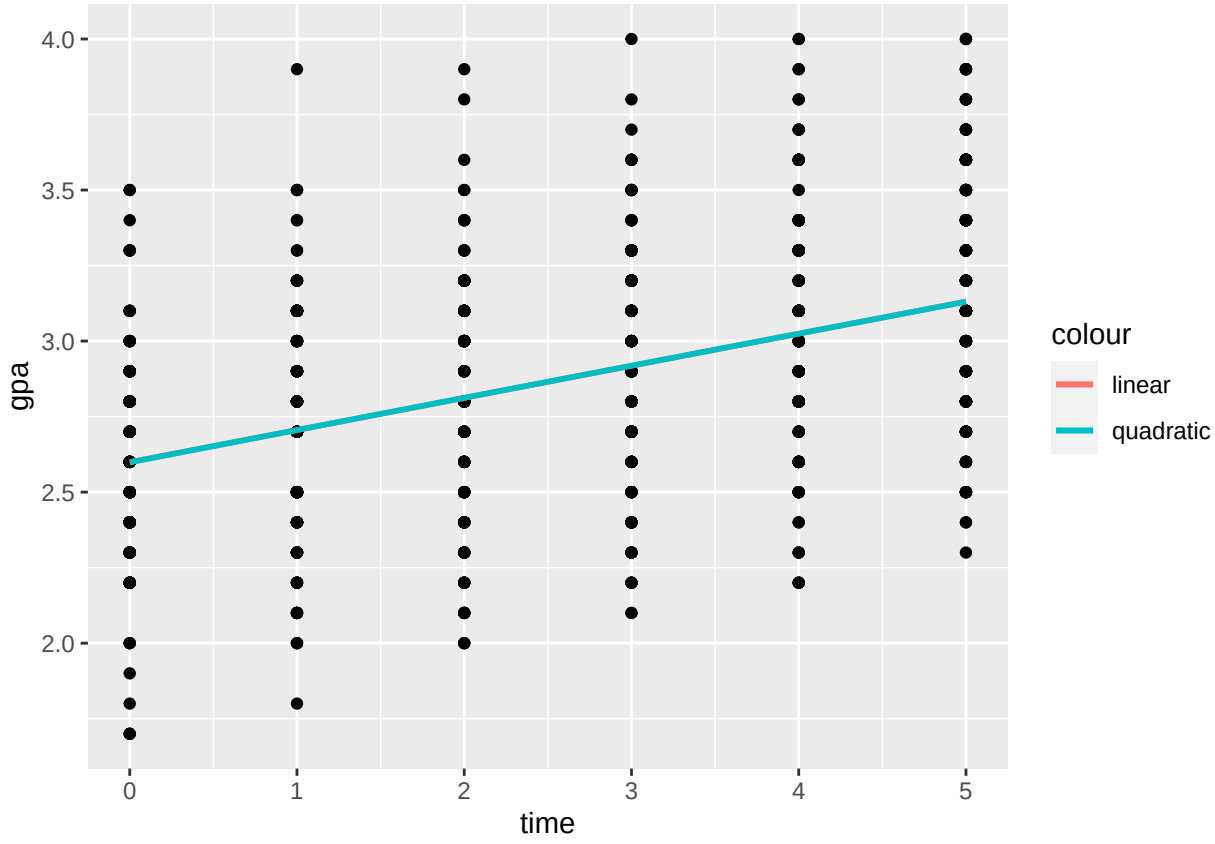
```
## # A tibble: 6 x 6
##   student   sex highgpa  time   gpa   job
##     <int> <dbl>   <dbl> <dbl> <dbl> <int>
## 1       1     1     2.8     0   2.3     2
## 2       1     1     2.8     1   2.1     2
## 3       1     1     2.8     2   3       2
## 4       1     1     2.8     3   3       2
## 5       1     1     2.8     4   3       2
## 6       1     1     2.8     5   3.3     2
```

5. Check the linearity assumption.

- Make a scatterplot of the variable GPA and Time (add a linear and quadratic fit line).

```
library(ggplot2)
ggplot(GPA_long,
aes(x = time, y = gpa)) +
        geom_point() +
           geom_smooth(method = "lm",
    aes(color = "linear"),
       se = FALSE) +
      geom_smooth(method = "lm",
     formula = y ~ x + I(x^2),
aes(color = "quadratic"),
se = FALSE)
```

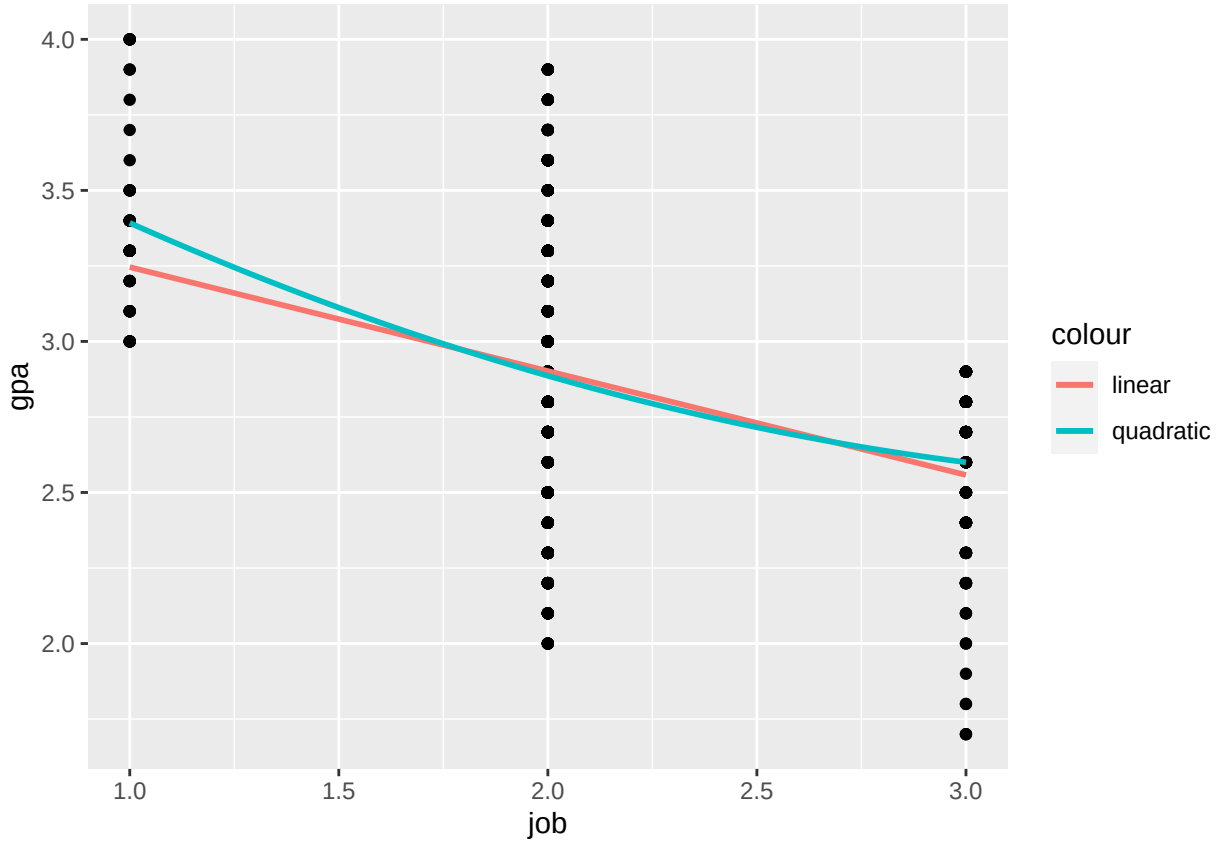
```
## `geom_smooth()` using formula 'y ~ x'
```



- Make a scatterplot of the variable GPA and Job (add a linear and quadratic fit line).

```
ggplot(GPA_long,
aes(x = job, y = gpa)) +
geom_point() +
geom_smooth(method = "lm",
aes(color = "linear"),
se = FALSE) +
geom_smooth(method = "lm",
formula = y ~ x + I(x^2),
aes(color = "quadratic"),
se = FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



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

Now we proceed with the multilevel analysis.

7. Specify the intercept only model.

• Inspect the output and compare the results with the model on the slides of the lecture. When they are the same, calculate and interpret the ICC.

```
library(lme4)
#recall from lab 1 that lmerTest will provide df, t, and p-values for fixed effects
library(lmerTest)
model_0 <- lm(gpa~1, data = GPA_long)
model_1 <- lmer(gpa~1 + (1|student) , REML = FALSE, data = GPA_long)
summary(model_0)
```

```
##
## Call:
## lm(formula = gpa ~ 1, data = GPA_long)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.165 -0.265 -0.065  0.235  1.135
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.86500    0.01135   252.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.393 on 1199 degrees of freedom
```

```
summary(model_1)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: gpa ~ 1 + (1 | student)
##    Data: GPA_long
##
##      AIC      BIC   logLik deviance df.resid
##    919.5    934.7   -456.7    913.5     1197
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.6505 -0.5505  0.0606  0.6353  2.5742
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  student  (Intercept) 0.05677  0.2383
##  Residual             0.09759  0.3124
## Number of obs: 1200, groups:  student, 200
##
## Fixed effects:
##              Estimate Std. Error        df t value Pr(>|t|)
## (Intercept)   2.86500    0.01911 199.99999   149.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model_1, model_0)
```

```
## Data: GPA_long
## Models:
## model_0: gpa ~ 1
## model_1: gpa ~ 1 + (1 | student)
##         npar     AIC     BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model_0    2 1167.28 1177.46 -581.64  1163.28
## model_1    3  919.46  934.73 -456.73   913.46 249.82  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ICC

```
#ICC <- var_occ/(var_occ+ var_sub)
ICC <-0.05677/(0.05677+0.09759)
```

8. Specify the model with Time as a linear predictor of the trend over time.

- Should you include the variable time uncentered or centered?
- Inspect the output and compare the results with the model on the slides of lecture. When they are the same, you can go on to the next step

```
model_2 <- lmer(gpa~1 + time + (1|student) , REML = FALSE, data = GPA_long)
summary(model_2)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: gpa ~ 1 + time + (1 | student)
##    Data: GPA_long
##
##      AIC      BIC   logLik deviance df.resid
##    401.6    422.0   -196.8    393.6     1196
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.6188 -0.6370 -0.0002  0.6366  2.8330
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  student  (Intercept) 0.06336  0.2517
##  Residual             0.05803  0.2409
## Number of obs: 1200, groups:  student, 200
##
## Fixed effects:
##              Estimate Std. Error        df t value Pr(>|t|)
## (Intercept) 2.599e+00  2.165e-02 3.244e+02  120.05   <2e-16 ***
## time        1.063e-01  4.072e-03 1.000e+03   26.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      (Intr)
## time -0.470
```

```
# to check if time is a significant predictor:
anova(model_2, model_1)
```

```
## Data: GPA_long
## Models:
## model_1: gpa ~ 1 + (1 | student)
## model_2: gpa ~ 1 + time + (1 | student)
##         npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model_1    3 919.46 934.73 -456.73   913.46
## model_2    4 401.65 422.01 -196.82   393.65 519.81  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

9. Set up the model with the time-varying predictor Job.

• Inspect and compare the results with the model on the slides of the lecture. When they are the same, calculate and interpret the explained variance at level 1 and level 2.

```
mean_job <- mean(GPA_long$job)
GPA_long$job <- GPA_long$job - mean_job
model_3 <- lmer(gpa~1 + time + job + (1|student) , REML = FALSE, data = GPA_long)
summary(model_3)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: gpa ~ 1 + time + job + (1 | student)
##    Data: GPA_long
##
##      AIC      BIC   logLik deviance df.resid
##    318.4    343.8   -154.2    308.4     1195
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0372 -0.6025 -0.0090  0.6451  2.8908
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  student  (Intercept) 0.05243  0.2290
##  Residual             0.05513  0.2348
## Number of obs: 1200, groups:  student, 200
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)    2.60883    0.02019  334.04217 129.227   <2e-16 ***
## time           0.10247    0.00399  995.04065  25.684   <2e-16 ***
## job           -0.17149    0.01811 1091.81282  -9.468   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      (Intr) time
## time -0.494
## job  -0.050  0.102
```

```
# to check if job is a significant predictor:
anova(model_2, model_3)
```

```
## Data: GPA_long
## Models:
## model_2: gpa ~ 1 + time + (1 | student)
## model_3: gpa ~ 1 + time + job + (1 | student)
##         npar    AIC    BIC  logLik deviance Chisq Df Pr(>Chisq)
## model_2    4 401.65 422.01 -196.82   393.65
## model_3    5 318.40 343.85 -154.20   308.40 85.25  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

10. Set up the model with the time-invariant predictors Gender and HighGPA.

• Inspect the output and compare the results with the model on the slides of the lecture. When they are the same, calculate and interpret the explained variance at level 2 and go on to the next step.

```
mean_highgpa <- mean(GPA_long$highgpa)
GPA_long$highgpa <- GPA_long$highgpa - mean_highgpa
model_4 <- lmer(gpa~1 + time + job + sex + highgpa + (1|student) , REML = FALSE,
data = GPA_long)
summary(model_4)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: gpa ~ 1 + time + job + sex + highgpa + (1 | student)
##    Data: GPA_long
##
##      AIC      BIC   logLik deviance df.resid
##    296.8    332.4   -141.4    282.8     1193
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.92840 -0.59427 -0.02739  0.63407  2.96127
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  student  (Intercept) 0.04497  0.2121
##  Residual             0.05514  0.2348
## Number of obs: 1200, groups:  student, 200
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)    2.53156    0.02593  266.99187  97.629  < 2e-16 ***
## time           0.10245    0.00399  994.90795  25.679  < 2e-16 ***
## job           -0.17221    0.01806 1100.57001  -9.534  < 2e-16 ***
## sex            0.14725    0.03305  194.13594   4.455 1.42e-05 ***
## highgpa        0.08470    0.02776  194.17579   3.051   0.0026 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##         (Intr) time   job    sex
## time    -0.387
## job     -0.057  0.102
## sex     -0.670  0.003  0.027
## highgpa -0.050  0.003  0.029  0.073
```

```
anova(model_3, model_4)
```

```
## Data: GPA_long
## Models:
## model_3: gpa ~ 1 + time + job + (1 | student)
## model_4: gpa ~ 1 + time + job + sex + highgpa + (1 | student)
##         npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model_3    5 318.40 343.85 -154.20   308.40
## model_4    7 296.76 332.39 -141.38   282.76 25.639  2  2.707e-06 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

11. For the time-varying predictors, check if the slopes are fixed or random.

• Start with the slope of Job: check if the variance of the slope for Job is significant. If the answer is no, you can remove u2 from the model and go on to the next step.

• Now turn to the variable Time: check if the variance of the slope for Time is significant. If the answer is yes, you can go on to the next step

```
#random slope for job
model_5a <- lmer(gpa~1 + time + job + sex + highgpa + (1+job|student) ,
REML = FALSE, data = GPA_long)
```

```
## boundary (singular) fit: see help('isSingular')
```

```
#summary(model_5a)
#anova(model_4, model_5a)
# random slope for time
model_5b <- lmer(gpa ~ 1 + time + job + sex + highgpa + (1 + time|student),
REML = F, data = GPA_long)
summary(model_5b)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: gpa ~ 1 + time + job + sex + highgpa + (1 + time | student)
##    Data: GPA_long
##
##      AIC      BIC   logLik deviance df.resid
##    188.1    233.9    -85.1    170.1     1191
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0175 -0.5379 -0.0002  0.5431  3.3656
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  student  (Intercept) 0.038233 0.19553
##           time        0.003837 0.06194  -0.21
##  Residual             0.041542 0.20382
## Number of obs: 1200, groups:  student, 200
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)  2.546e+00  2.391e-02  2.205e+02 106.458  < 2e-16 ***
## time         1.034e-01  5.586e-03  1.996e+02  18.506  < 2e-16 ***
## job         -1.311e-01  1.726e-02  1.038e+03  -7.595 6.84e-14 ***
## sex          1.157e-01  3.130e-02  1.978e+02   3.696 0.000284 ***
## highgpa      8.854e-02  2.628e-02  1.976e+02   3.369 0.000907 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Correlation of Fixed Effects:
##         (Intr) time   job    sex
## time    -0.320
## job     -0.061  0.069
## sex     -0.688  0.002  0.030
## highgpa -0.051  0.002  0.022  0.073
```

```
#anova(model_5b, model_4)
```

12. Now, we'll check if Gender can (partly) explain why the trajectory over time differs between students. That is, we include Gender as a predictor of the slope for Time.

• Inspect the output and compare the results with the models on the slides of the lecture. When they are the same, calculate and interpret the explained slope variance.

• Make an interaction plot to show how the trajectory over time differs between boys and girls.

• For R one option is to draw an empty plot with time on the x-axis and GPA on the y-axis. Next, draw in the appropriate lines using abline() and the corresponding equations for the predicted outcomes over time for boys and girls.

```
model_6 <- lmer(gpa ~ 1 + time + job + sex + highgpa + sex * time + (1 + time|student),
REML = F, data = GPA_long)
summary(model_6)
```

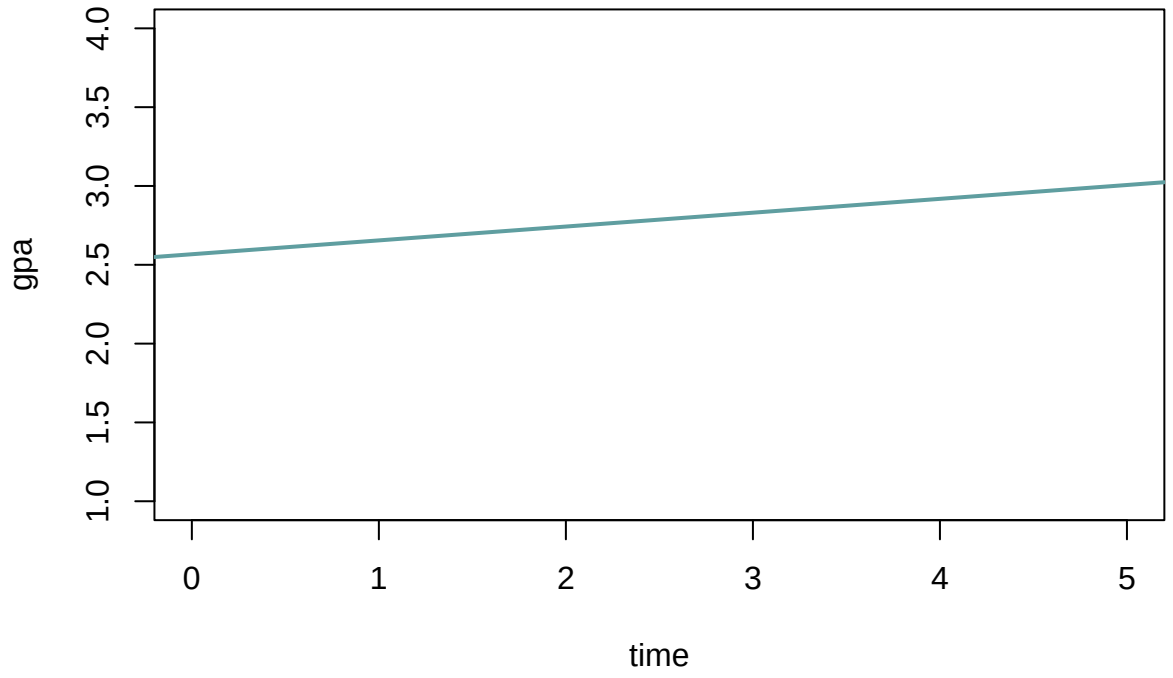
```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: gpa ~ 1 + time + job + sex + highgpa + sex * time + (1 + time |
##     student)
##    Data: GPA_long
##
##      AIC      BIC   logLik deviance df.resid
##    183.0    233.9    -81.5    163.0     1190
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0044 -0.5268 -0.0138  0.5290  3.3513
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  student  (Intercept) 0.037811 0.19445
##           time        0.003614 0.06012  -0.19
##  Residual             0.041555 0.20385
## Number of obs: 1200, groups:  student, 200
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)  2.567e+00  2.512e-02  2.012e+02 102.203  < 2e-16 ***
## time         8.783e-02  7.951e-03  1.987e+02  11.046  < 2e-16 ***
## job         -1.321e-01  1.723e-02  1.042e+03  -7.670 3.93e-14 ***
## sex          7.551e-02  3.465e-02  2.003e+02   2.179 0.030499 *
## highgpa      8.850e-02  2.627e-02  1.976e+02   3.369 0.000907 ***
## time:sex     2.956e-02  1.096e-02  1.977e+02   2.698 0.007581 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr) time   job    sex    highgp
## time     -0.432
## job      -0.060  0.054
## sex      -0.725  0.312  0.030
## highgpa  -0.049  0.001  0.022  0.066
## time:sex  0.312 -0.724 -0.008 -0.430  0.000
```

```
anova(model_6, model_5b)
```

```
## Data: GPA_long
## Models:
## model_5b: gpa ~ 1 + time + job + sex + highgpa + (1 + time | student)
## model_6: gpa ~ 1 + time + job + sex + highgpa + sex * time + (1 + time | student)
##          npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model_5b    9 188.12 233.93 -85.059   170.12
## model_6    10 182.97 233.87 -81.486   162.97 7.1464  1   0.007511 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# interaction plot
time <- 0:5
plot(x = time, ylim = c(1,4), xlim = c(0,5), type = "n", xlab = "time", ylab = "gpa")
col.sex <- c("cadetblue", "coral")
abline(a = model_6@beta[1], b = model_6@beta[2], col = col.sex[1], lwd = 2)
```

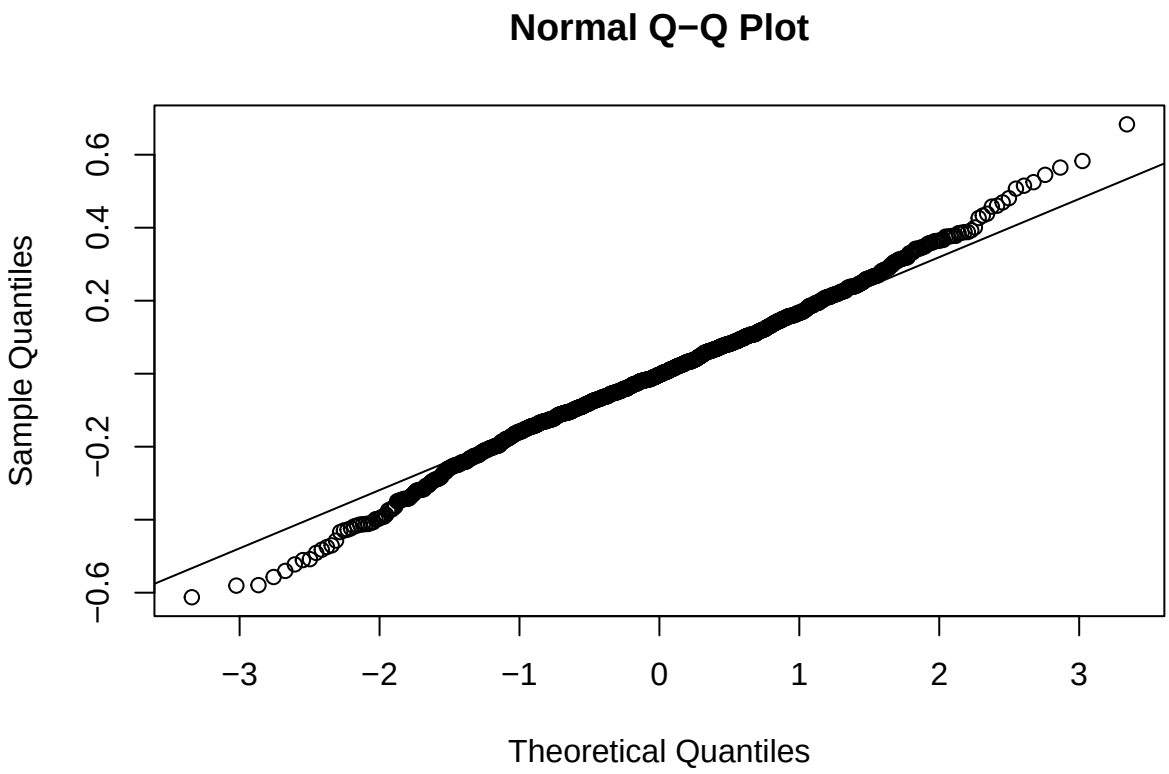



```
#girls start out higher and proceed faster over time:
#abline(a = model_6@beta[1] + model_6@beta[4], b = model_6@beta[2] + model_6@beta[6],
#col = col.sex[2], lwd = 2)
#legend("topleft", bty = "n", lwd = 2, col = rev(col.sex),
#legend = c("girls", "boys"))
```

13. Check the normality assumption for the level 1 residuals.

- The level 1 and 2 residuals can be accessed via the function residuals(object, . . .). Check the help file for more information and note the argument level and test what it does!
- Using the obtained residuals, (visually) inspect the normality assumption.
- Are the assumptions met?- yes

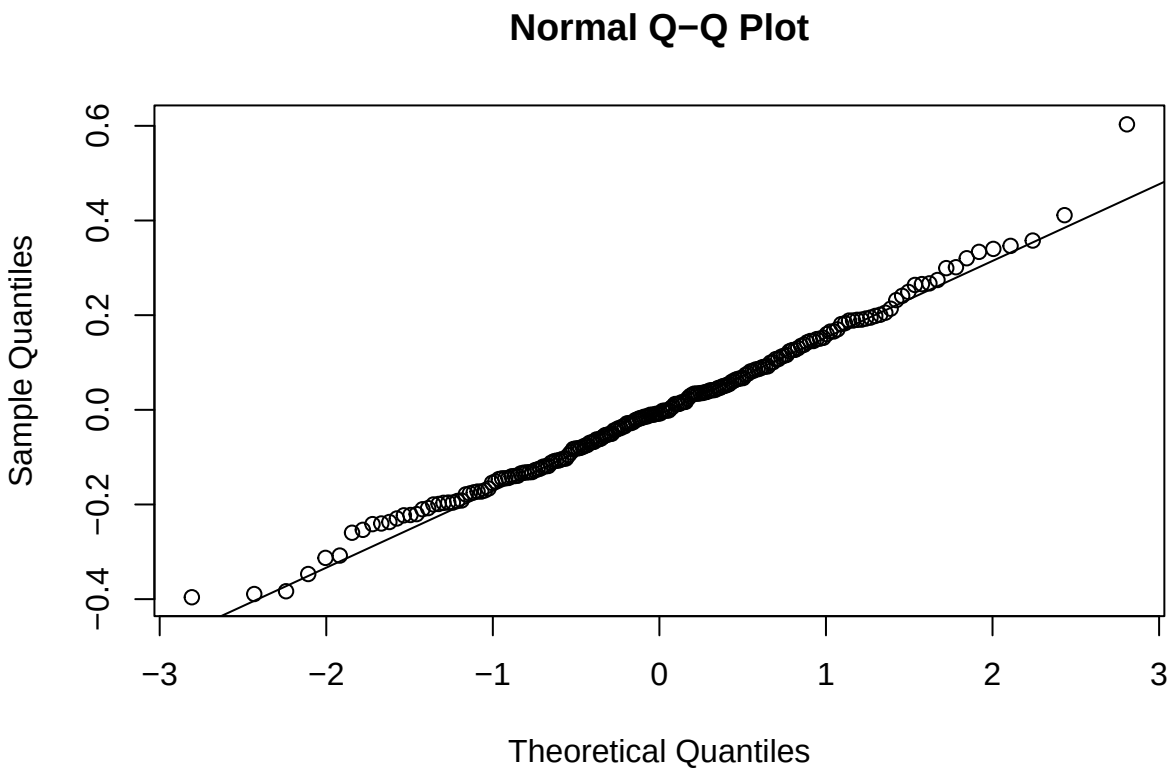
```
qqnorm(residuals(model_6))
qqline(residuals(model_6))
```



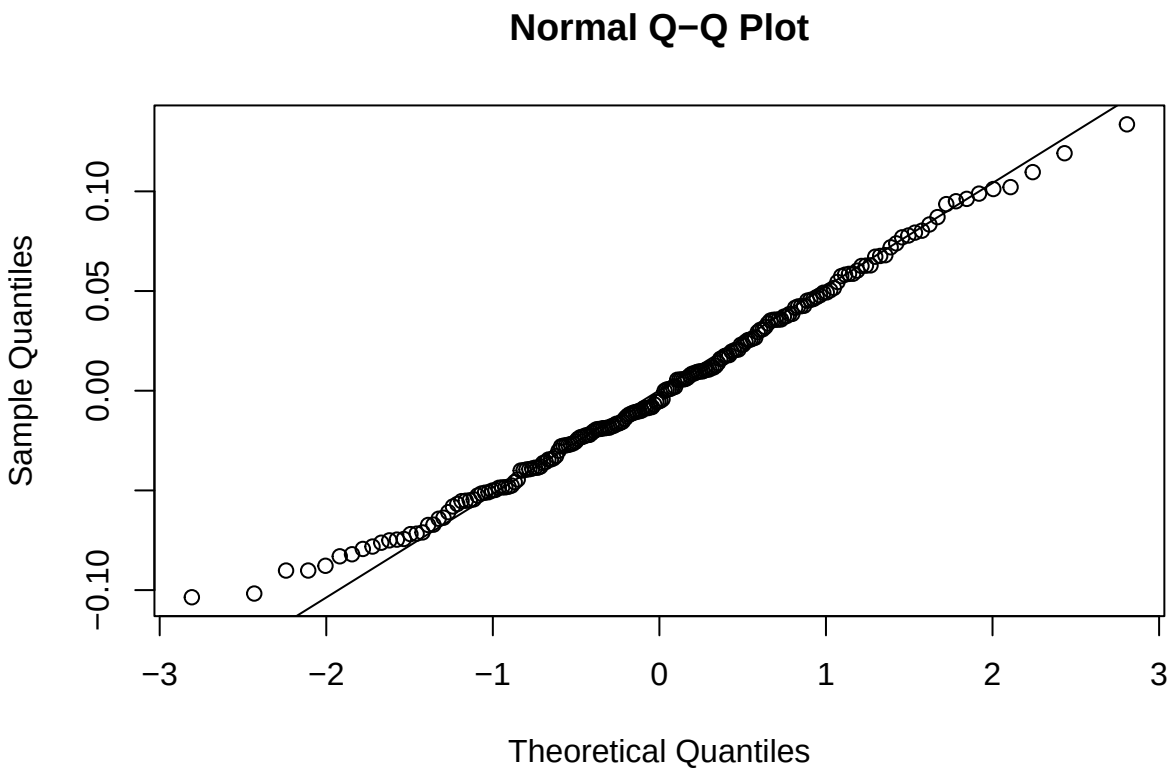
14. Check the normality assumption for the level 2 residuals

- Are the assumptions met for the intercept and slope errors?

```
#intercept
qqnorm(ranef(model_6)$student[,1])
qqline(ranef(model_6)$student[,1])
```




```
#slope
qqnorm(ranef(model_6)$student[,2])
qqline(ranef(model_6)$student[,2])
```


Normal Q−Q Plot
Sample Quantiles
−0.10
0.00
0.05
0.10
−3
−2
−1
0
1
2
3
Theoretical Quantiles