Problem Set 6: Longitudinal wtsa and Three-Level Models

S-043 (Fall 2023)

```
## Rows: 198
## Columns: 6
## $ id     <dbl> 45, 45, 45, 45, 45, 258, 258, 258, 258, 287, 287, 287, 287, 483~
## $ occ    <dbl> 1, 2, 3, 4, 5, 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 1, 2, 1, 2, 1, ~
## $ age    <dbl> 0.137, 0.657, 1.218, 1.429, 2.272, 0.192, 0.687, 1.128, 2.305, ~
## $ weight <dbl> 5.17, 10.86, 13.15, 13.20, 15.88, 5.30, 9.74, 9.98, 11.34, 4.82~
## $ brthwt <dbl> 4140, 4140, 4140, 4140, 4140, 3155, 3155, 3155, 3155, 3850, 385~
## $ gender <dbl> 1, 1, 1, 1, 1, 2, 2, 2, 2, 1, 1, 1, 1, 2, 2, 2, 2, 2, 1, 1, 1, ~
```

```
##        id            occ             age            weight          brthwt
##  Min.   :  45   Min.   :1.00   Min.   :0.115   Min.   : 3.10   Min.   :1575
##  1st Qu.:1271   1st Qu.:1.00   1st Qu.:0.646   1st Qu.: 7.01   1st Qu.:2807
##  Median :2351   Median :2.00   Median :0.997   Median : 9.07   Median :3120
##  Mean   :2491   Mean   :2.11   Mean   :1.081   Mean   : 8.84   Mean   :3106
##  3rd Qu.:3704   3rd Qu.:3.00   3rd Qu.:1.471   3rd Qu.:10.89   3rd Qu.:3390
##  Max.   :4975   Max.   :5.00   Max.   :2.546   Max.   :17.20   Max.   :4270
##      gender
##  Min.   :1.0
##  1st Qu.:1.0
##  Median :1.0
##  Mean   :1.5
##  3rd Qu.:2.0
##  Max.   :2.0
```

Linear and Quadratic Growth Models

1a. Plot

No need to turn in anything here.

1b. Fit a Linear Growth Model to these Data

```
##center the data
wts$age_center = wts$age - mean(wts$age)

m1 <- lmer(weight ~ age_center + gender + (1 + age_center |id), data = wts)

summary(m1)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
```

```
## lmerModLmerTest]
## Formula: weight ~ age_center + gender + (1 + age_center | id)
##    Data: wts
##
## REML criterion at convergence: 684
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.3842 -0.6958 -0.0423  0.7348  2.1369
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  id       (Intercept) 0.534    0.731
##           age_center  0.207    0.455    1.00
##  Residual             1.370    1.170
## Number of obs: 198, groups:  id, 68
##
## Fixed effects:
##             Estimate Std. Error     df t value Pr(>|t|)
## (Intercept)    9.803      0.368 72.256   26.62   <2e-16 ***
## age_center     3.459      0.126 70.931   27.36   <2e-16 ***
## gender        -0.635      0.232 71.316   -2.74   0.0077 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##            (Intr) ag_cnt
## age_center  0.127
## gender     -0.942 -0.006
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

1c. Fit a Quadratic Growth Model

The model fails because the number of observations is less than the number of random effect parameters. By review of the histogram we can visualize the distribution of number of observations per child, the majority of children have 3 or 4 observations, but some have 5 and some have as few as 1 or 2.

```
##squared term
wts$age_sq = wts$age_center^2

#the model below will not run
#m2 <- lmer(weight ~ age_center + age_sq + gender + (1 + age_center + age_sq|id), data = wts)

table(wts$id) # tally up obs/child
```

```
##
##   45  258  287  483  725  800  801  832  833  944 1005 1010 1055 1141 1155 1249
##    5    4    4    3    2    2    2    3    2    3    2    3    4    4    3    1
## 1264 1291 1341 1417 1441 1572 1614 1976 2088 2106 2108 2130 2150 2197 2251 2302
##    3    3    3    3    3    3    3    4    3    2    4    3    3    3    2    2
## 2317 2351 2353 2433 2615 2661 2730 2817 2850 3021 3148 3171 3192 3289 3446 3556
##    2    4    3    4    3    3    2    4    3    3    4    2    3    4    4    3
```

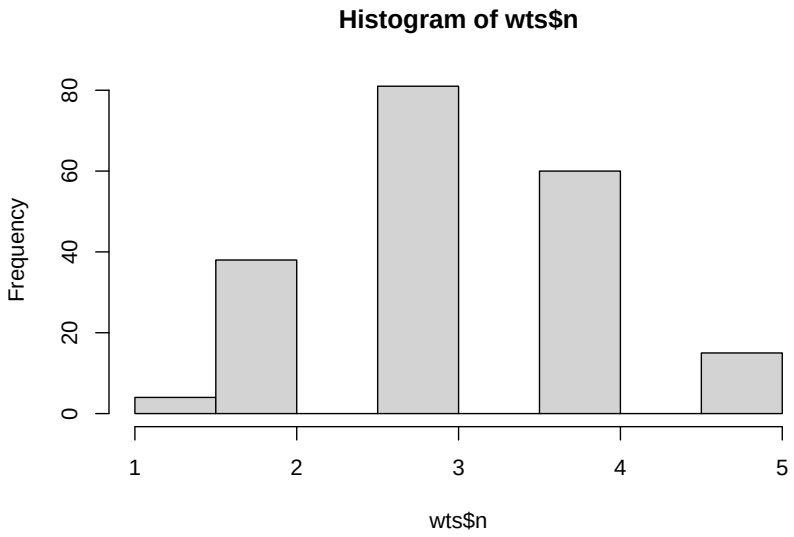
```
## 3672 3714 3827 3860 4017 4081 4108 4119 4139 4152 4169 4357 4418 4475 4518 4666
##    3    1    5    3    1    1    3    3    4    4    2    2    2    2    2    2
## 4682 4826 4841 4975
##    5    4    2    2
```

```
table(table(wts$id)) # tally the tallys
```

```
##
##  1  2  3  4  5
##  4 19 27 15  3
```

```
# Add number of obs for each group to your data.
wts = wts %>% group_by(id) %>%
mutate(n = n())

hist(wts$n)
```



1d. Subset the Data to fit the Quadratic Model

We have subset the data to remove students who had only 1 observation, this allowed the quadratic model to converge. In terms of data loss, 4 participants (each with 1 observation so 4 unique observations), were removed.

```
subset <- filter(wts, n != 1)

subset$age_center = subset$age - mean(subset$age)
subset$age_sq = subset$age_center^2

m3 <- lmer(weight ~ age_center + age_sq + gender + (1 + age_center + age_sq|id), data = subset)

summary(m3)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: weight ~ age_center + age_sq + gender + (1 + age_center + age_sq |
##     id)
##    Data: subset
##
## REML criterion at convergence: 481
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1269 -0.3953  0.0038  0.3726  2.5642
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  id       (Intercept) 1.207    1.099
##           age_center  0.301    0.548     0.79
##           age_sq      0.269    0.519    -0.74 -0.20
##  Residual             0.221    0.470
## Number of obs: 194, groups:  id, 64
##
## Fixed effects:
##             Estimate Std. Error     df t value Pr(>|t|)
## (Intercept)   10.580      0.290 76.251   36.43   <2e-16 ***
## age_center     4.078      0.090 57.899   45.30   <2e-16 ***
## age_sq        -1.686      0.103 45.694  -16.34   <2e-16 ***
## gender        -0.441      0.164 51.781   -2.69   0.0097 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##            (Intr) ag_cnt age_sq
## age_center  0.347
## age_sq     -0.366 -0.258
## gender     -0.851 -0.003  0.010
```

1e. Linear vs. Quadratic

For the first comparison, we will work on the same subset dataset for both models (where those with only one observation were removed).

Using a likelihood ratio test (calculated through ANOVA), we find that the quadratic model is a better fit for the data.

We also want to test whether the linear model including all the data (so not subset to omit those with only 1 observation), is a better fit than the quadratic model. Per the R2, the quadratic model is a better fit (0.976 vs 0.856); however, R2 nearly always increases with a more complex model, and the slight increase in R2 may not be worth the loss of data from 4 students.

```
# linear model on subset data
m1_subset <- lmer(weight ~ age_center + gender + (1 + age_center|id), data = subset)
summary(m1_subset)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
```

```
## Formula: weight ~ age_center + gender + (1 + age_center | id)
##    Data: subset
##
## REML criterion at convergence: 668
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.3945 -0.6883 -0.0546  0.7446  2.1615
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  id       (Intercept) 0.530    0.728
##           age_center  0.211    0.459    1.00
##  Residual             1.353    1.163
## Number of obs: 194, groups:  id, 64
##
## Fixed effects:
##             Estimate Std. Error     df t value Pr(>|t|)
## (Intercept)    9.819      0.371 70.152   26.47   <2e-16 ***
## age_center     3.450      0.126 70.305   27.28   <2e-16 ***
## gender        -0.642      0.233 69.366   -2.75   0.0076 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##            (Intr) ag_cnt
## age_center  0.122
## gender     -0.941  0.001
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

```
## quadratic model again on subset data (same as q1d)
m3 <- lmer(weight ~ age_center + age_sq + gender + (1 + age_center + age_sq|id), data = subset)
summary(m3)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: weight ~ age_center + age_sq + gender + (1 + age_center + age_sq |
##     id)
##    Data: subset
##
## REML criterion at convergence: 481
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1269 -0.3953  0.0038  0.3726  2.5642
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  id       (Intercept) 1.207    1.099
##           age_center  0.301    0.548     0.79
##           age_sq      0.269    0.519    -0.74 -0.20
##  Residual             0.221    0.470
## Number of obs: 194, groups:  id, 64
```

```
##
## Fixed effects:
##             Estimate Std. Error     df t value Pr(>|t|)
## (Intercept)   10.580      0.290 76.251   36.43   <2e-16 ***
## age_center     4.078      0.090 57.899   45.30   <2e-16 ***
## age_sq        -1.686      0.103 45.694  -16.34   <2e-16 ***
## gender        -0.441      0.164 51.781   -2.69   0.0097 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##            (Intr) ag_cnt age_sq
## age_center  0.347
## age_sq     -0.366 -0.258
## gender     -0.851 -0.003  0.010
```

```
anova(m1_subset, m3)
```

```
## Data: subset
## Models:
## m1_subset: weight ~ age_center + gender + (1 + age_center | id)
## m3: weight ~ age_center + age_sq + gender + (1 + age_center + age_sq | id)
##           npar AIC BIC logLik deviance Chisq Df Pr(>Chisq)
## m1_subset    7 676 699   -331      662
## m3          11 492 528   -235      470   192  4     <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
### now test our original linear growth model with ALL the data
m1 <- lmer(weight ~ age_center + gender + (1 + age_center |id), data = wts)
summary(m1)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: weight ~ age_center + gender + (1 + age_center | id)
##    Data: wts
##
## REML criterion at convergence: 684
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.3842 -0.6958 -0.0423  0.7348  2.1369
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  id       (Intercept) 0.534    0.731
##           age_center  0.207    0.455    1.00
##  Residual             1.370    1.170
## Number of obs: 198, groups:  id, 68
##
## Fixed effects:
##             Estimate Std. Error     df t value Pr(>|t|)
## (Intercept)    9.803      0.368 72.256   26.62   <2e-16 ***
```

```
## age_center     3.459      0.126 70.931   27.36   <2e-16 ***
## gender        -0.635      0.232 71.316   -2.74   0.0077 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) ag_cnt
## age_center  0.127
## gender     -0.942 -0.006
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

```
##compare it to our quadratic model. We can't use LRT because the datasets are different. We will use R
# install.packages("MuMIn")
library(MuMIn)

r.squaredGLMM(m1)
```

```
##        R2m   R2c
## [1,] 0.787 0.856
```

```
r.squaredGLMM(m3)
```

```
##        R2m   R2c
## [1,] 0.863 0.976
```

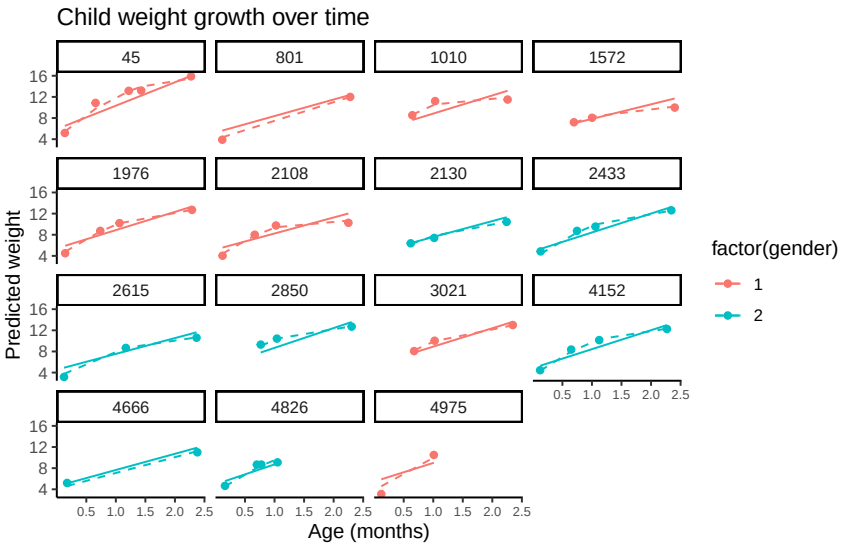
1f. Plot the Growth Curves

```
library(ggplot2)

selected_ids <- sample(subset$id, 16)

# Filter the dataset to include only the selected children
selected_wts <- subset[subset$id %in% selected_ids, ]

ggplot(data = selected_wts, aes(x = age, y = weight, group = id, color = factor(gender))) +
  facet_wrap(~ id) +
  geom_point() +
  geom_line(aes(y = predict(m3, newdata = selected_wts), color = factor(gender)), linetype = "dashed") +
 labs(title = "Child weight growth over time",
x = "Age (months)",
y = "Predicted weight",
caption = "Age is centered at 1.08 years.
Dashed lines are for the quadratic growth model and solid lines are for linear growth model") +
theme(axis.text.x = element_text(size = rel(.75), angle = 00))
```



Age is centered at 1.08 years.
Dashed lines are for the quadratic growth model and solid lines are for linear growth model

Piecewise Linear Growth (READS wtsa)

2a. Mathematical Model

$$
\begin{aligned}
score_{tjk} &= \beta_{0jk} + \beta_{1jk} school_{tjk} + \beta_{2jk} summer_{tjk} + \varepsilon_{tjk} \\
\beta_{0jk} &= \gamma_{00k} + \mu_{0jk} \\
\beta_{1jk} &= \gamma_{10k} + \mu_{1jk} \\
\beta_{2jk} &= \gamma_{20k} + \mu_{2jk} \\
\gamma_{00k} &= \alpha_{000} + \nu_{0k} \\
\gamma_{10k} &= \alpha_{100} + \nu_{1k} \\
\gamma_{20k} &= \alpha_{200} + \nu_{2k}
\end{aligned}
$$

where t is time j is a given student and k is a given school

where $score_{ijk}$ is the score for the jth student in the kth school at the t time point.

β_{0jk} and β_{1jk} and $beta_2jk$ is the grand intercept or average across all schools, as well as the average learning rate during school months and for summer months respectively.

ε_{ijk} represents the (within-person) residual error at the individual observation level.

μ_{0jk} and μ_{1jk} and μ_{2jk} represent random effects at the student level, or how student j differs from the average student in school k.

γ_{00k} and γ_{10k} and γ_{20k} represent average student intercept and growth rates (for school months and summer months respectively) in school k.

ν_{0k} and ν_{1k} and ν_{2k} represent random effects at the school level or rather how school k differs from the average school with respect to the random intercept of the school, or random effect of school and summer months on learning rates respectively.

2b. Answer RQ1: Do students learn more quickly over the summer or school year?

```
studs <- readRDS("~/Desktop/Fall 2023/Multilevel-HW/PS 6 /READS_student_data.Rds")

studs$sum2=studs$summer-studs$school


m2 <- lmer(score ~ 1 + sum2 + (1+sum2|id) + (1+sum2|sch), data = studs)
display(m2)
```

```
## lmer(formula = score ~ 1 + sum2 + (1 + sum2 | id) + (1 + sum2 |
##     sch), data = studs)
##             coef.est coef.se
## (Intercept) 185.75     1.37
## sum2         -1.71     0.06
##
## Error terms:
##  Groups   Name        Std.Dev. Corr
##  id       (Intercept) 20.68
##           sum2         0.45    -1.00
##  sch      (Intercept)  7.54
```

```
##           sum2         0.19    -0.11
##  Residual             12.63
## ---
## number of obs: 5000, groups: id, 1250; sch, 40
## AIC = 42915.3, DIC = 42894
## deviance = 42895.8
```

```
confint(m2)
```

```
##                2.5 %  97.5 %
## .sig01        19.736  21.686
## .sig02        -1.000   1.000
## .sig03         0.357   0.599
## .sig04         5.535   9.939
## .sig05        -1.000   1.000
## .sig06         0.000   0.326
## .sigma        12.345  12.920
## (Intercept) 183.041 188.461
## sum2          -1.816  -1.594
```

The model predicts an increase of 0.19 units in the 'score' outcome variable for a one unit increase in summer months to the baseline learning growth rate. This is statistically significant with a 95% CI of [-1.82, -1.60] that does not contain the null value.

The level 1 model for this lmer code is

$$score_{tjk} = \beta_{0jk} + \beta_{1jk} * summer_{tjk} + \varepsilon_{tjk}$$

where summer is summer months minus school months.

2c. Answer RQ2: Do student's rates of growth depend on student level poverty?

```
m_3 <- lmer(score ~ 1 + sum2*frl + (1 + sum2|id) + (1 + sum2*frl|sch),
                  data = studs)
display(m_3)
```

```
## lmer(formula = score ~ 1 + sum2 * frl + (1 + sum2 | id) + (1 +
##     sum2 * frl | sch), data = studs)
##             coef.est coef.se
## (Intercept) 195.21     2.15
## sum2         -1.92     0.11
## frl         -13.16     1.97
## sum2:frl      0.27     0.13
##
## Error terms:
##  Groups   Name        Std.Dev. Corr
##  id       (Intercept) 19.75
##           sum2         0.45    -1.00
##  sch      (Intercept) 10.21
##           sum2         0.27     0.55
##           frl          7.25    -0.88 -0.78
```

```
##           sum2:frl     0.36    -0.32 -0.80  0.36
##  Residual             12.61
## ---
## number of obs: 5000, groups: id, 1250; sch, 40
## AIC = 42827.3, DIC = 42789
## deviance = 42790.1
```

Note that we are assuming that free and reduced lunch eligibility for a student j may vary in across schools, but not within schools (as indicated by the inclusion of free or reduced lunch variable into random slope for schools but NOT for individuals)

The positive value of 0.37 means that there is between school variability in the interaction effect between free and reduced lunch and the effect of summer (months) on the learning scores for students In other words, the influence of the interaction between free or reduced lunch eligibility (frl) and the difference variable summer 2 on student scores differs across schools. This difference is above what is explained by the fixed effects interaction in the model, as well as the fixed effects of frl and summer months (time) in the variation of student scores.

2d. Concept Check 1

It would be theoretically possible but we would run into several issues with this more complex model that may be unnecessary. We would more than likely have issues with multidisciplinary in the model, given the imbalance between some schools having free or reduced lunch eligibility for ALL of their students and some schools not having any students who are not eligible reducing the stability of our model, and leading to loss of precision potentially. With respect to modeling a random slope which is intent on showcasing the variability between schools (i.e how much different a school is from another school, and how this variability impacts the outcome of student reading scores leading to potential convergence issues as we do not have enough observations to estimate meaningful random effects within the higher unit (school level).