HW 4

Iris Zhong

2022-11-18

library(nlme)  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✔ ggplot2 3.3.5 ✔ purrr 0.3.4  
## ✔ tibble 3.1.6 ✔ dplyr 1.0.8  
## ✔ tidyr 1.2.0 ✔ stringr 1.4.0  
## ✔ readr 2.1.2 ✔ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::collapse() masks nlme::collapse()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(data.table)

##   
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

## The following object is masked from 'package:purrr':  
##   
## transpose

library(lubridate)

##   
## Attaching package: 'lubridate'

## The following objects are masked from 'package:data.table':  
##   
## hour, isoweek, mday, minute, month, quarter, second, wday, week,  
## yday, year

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

# Research question

As seen from the exploration last time, no growth over time can be observed from monthly aggregated assessment scores. So in this assignment, I’m going to investigate whether intrapersonal variances can be explained by any time-invariant covariate. In particular, ASSISTments, the digital learning platform where I retrieved the data, has designed affect detectors in their systems. How much a student is concentrated, frustrated, bored, or confused is measured on a 0-1 continuous scale. For this assignment, I will use the frustration detector to explore: Does a student’s overall frustration when answering questions in ASSISTments influence their variance in monthly average assessment score? This is an objective 5 question: analysis of determinats of interindividual differences in intraindividual change.

My hypothesis is that students who have a high frustration score might answer questions wrong relatively frequently. Thus, their scores are consistently low, and the fluctuation in their scores is small. On the other hand, students who don’t feel as frustruated should have higher average scores, but they can still answer questions wrong sometimes due to careless mistakes, so their scores might exhibit larger variances.

# Data cleaning

# read data  
data <- read\_csv("processed\_data.csv")

## New names:  
## Rows: 6123270 Columns: 30  
## ── Column specification  
## ──────────────────────────────────────────────────────── Delimiter: "," chr  
## (4): skill, problem\_type, tutor\_mode, type dbl (24): ...1, problem\_log\_id,  
## problem\_id, user\_id, assignment\_id, assistm... dttm (2): start\_time, end\_time  
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ  
## Specify the column types or set `show\_col\_types = FALSE` to quiet this message.  
## • `` -> `...1`

data\_by\_month <- data %>%   
 group\_by(user\_id, month = month(start\_time)) %>%   
 summarize(month\_avg = mean(correct),  
 assignment\_count\_per\_month = uniqueN(assignment\_id),  
 problem\_count\_per\_month = uniqueN(problem\_id))

## `summarise()` has grouped output by 'user\_id'. You can override using the  
## `.groups` argument.

keep persons that have responses in at least 6 months

data\_by\_month <- data\_by\_month %>%  
 group\_by(user\_id) %>%  
 mutate(n\_month = n()) %>%  
 filter(n\_month > 6) %>%  
 ungroup()

keep persons that have tried at least 10 items each month they participated

data\_by\_month <- data\_by\_month %>%  
 mutate(item\_count\_per\_month\_over\_10 = ifelse(problem\_count\_per\_month >= 10, 1, 0)) %>%  
 group\_by(user\_id) %>%  
 mutate(item\_count\_per\_month\_over\_10 = mean(item\_count\_per\_month\_over\_10)) %>%  
 filter(item\_count\_per\_month\_over\_10 == 1) %>%  
 select(-item\_count\_per\_month\_over\_10) %>%  
 ungroup()

Add number of items/assignments by learner

Add the average score of frustration, concentrated, boredom for each person (each item attempt is marked with these affect scores)

data\_by\_month <- data\_by\_month %>%  
 group\_by(user\_id) %>%  
 # overall number of items/assignments done by a learner   
 mutate(n\_problem = sum(problem\_count\_per\_month),  
 n\_assignment = sum(assignment\_count\_per\_month)) %>%  
 # mean affect scores for an individual  
 left\_join(data %>% group\_by(user\_id) %>% summarize(mean\_frustrated = mean(frustrated\_confidence),   
 mean\_concentrated = mean(concentrating\_confidence),  
 mean\_bored = mean(bored\_confidence)), by = "user\_id")

Centering the frustration, concentration, and boredom variables

# find the means of these variables (each person is counted once)  
psych::describe(data\_by\_month %>% distinct(user\_id, mean\_frustrated, mean\_concentrated, mean\_bored))

## vars n mean sd median trimmed mad  
## user\_id 1 2653 165529.42 27129.54 176639.00 170157.72 7908.19  
## mean\_frustrated 2 2653 0.39 0.01 0.39 0.39 0.01  
## mean\_concentrated 3 2653 0.73 0.03 0.74 0.73 0.02  
## mean\_bored 4 2653 0.25 0.04 0.26 0.25 0.04  
## min max range skew kurtosis se  
## user\_id 74601.00 200646.00 126045.00 -1.49 0.90 526.71  
## mean\_frustrated 0.36 0.43 0.07 0.50 0.80 0.00  
## mean\_concentrated 0.49 0.76 0.28 -2.30 11.48 0.00  
## mean\_bored 0.10 0.37 0.28 -0.27 0.13 0.00

# Subtract mean from individual scores  
data\_by\_month <- data\_by\_month %>%  
 mutate(mean\_frustratedC = mean\_frustrated - 0.39,  
 mean\_concentratedC = mean\_concentrated - 0.73,  
 mean\_boredC = mean\_bored - 0.25)

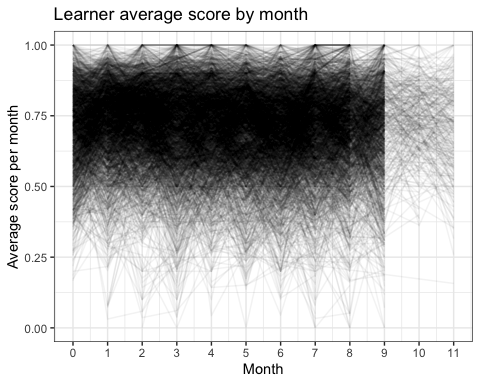
Recode month so that the month number reflects the number of months passed since September. (September coded as 0, October as 1, …, August as 11)

data\_by\_month <- data\_by\_month %>%  
 mutate(month = case\_when(month > 8 ~ month - 9,  
 month <= 8 ~ month + 3))

# Visualizations from raw data

This is the graph from HW 3.

ggplot(data = data\_by\_month, #data set  
 aes(x = month, y = month\_avg, group = user\_id)) + #setting variables  
 #geom\_point(size=.5, alpha = 0.01) + #adding points to plot  
 geom\_line(alpha = 0.05) + #adding lines to plot  
 theme\_bw() + #changing style/background  
 scale\_x\_continuous(breaks = seq(0, 11)) +  
 labs(x = "Month",  
 y = "Average score per month",  
 title = "Learner average score by month")

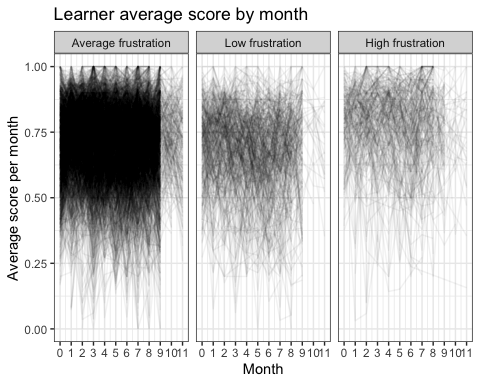


No trend of growth can be observed from the figure.

Color the plot by their frustrated score: greater than +1SD, between -1 and +1 SD, smaller than -1SD.

# Construct group based on SD  
frustrated\_plot <- data\_by\_month %>%  
 mutate(frustrated\_group = case\_when(mean\_frustratedC > 0.01 ~ "High frustration",  
 mean\_frustratedC < -0.01 ~ "Low frustration",  
 T ~ "Average frustration")) %>%  
 mutate(frustrated\_group = as.factor(frustrated\_group))

ggplot(data = frustrated\_plot, #data set  
 aes(x = month, y = month\_avg, group = user\_id)) + #setting variables  
 #geom\_point(size=.5, alpha = 0.01) + #adding points to plot  
 geom\_line(alpha = 0.05) + #adding lines to plot  
 theme\_bw() + #changing style/background  
 scale\_x\_continuous(breaks = seq(0, 11)) +  
 labs(x = "Month",  
 y = "Average score per month",  
 title = "Learner average score by month") +  
 facet\_wrap(~frustrated\_group)



Interestingly, high frustration students seem to on average perform better than low frustration students. Possibly, high-performing students become frustrated when constantly given easy questions. The difference of variance among groups is not easily observed.

# Model specification

First, I will fit a model assuming homogeneous intrapersonal variance, but controlling for frustration in fixed effect.

.

The next model allows within-person variance to vary by the magnitude of frustration.

, and .

# Model fitting

## Model 1

model\_1 = lme(fixed = month\_avg ~ 1 + mean\_frustratedC,  
 random = list(user\_id = pdSymm(form = ~ 1)),  
 data = data\_by\_month,  
 na.action = na.exclude,  
 method = 'REML')  
summary(model\_1)

## Linear mixed-effects model fit by REML  
## Data: data\_by\_month   
## AIC BIC logLik  
## -29591.12 -29558.82 14799.56  
##   
## Random effects:  
## Formula: ~1 | user\_id  
## (Intercept) Residual  
## StdDev: 0.1028169 0.115651  
##   
## Fixed effects: month\_avg ~ 1 + mean\_frustratedC   
## Value Std.Error DF t-value p-value  
## (Intercept) 0.7186822 0.00216668 21144 331.6981 0  
## mean\_frustratedC 3.1569783 0.26541504 2651 11.8945 0  
## Correlation:   
## (Intr)  
## mean\_frustratedC 0.172   
##   
## Standardized Within-Group Residuals:  
## Min Q1 Med Q3 Max   
## -5.47969735 -0.51507243 0.07573286 0.59073088 4.27365938   
##   
## Number of Observations: 23797  
## Number of Groups: 2653

The fixed effect suggests that, a learner with an average level of frustration is predicted to have a monthly average score of 0.72. If frustration increases by 0.1, the predicted score is 0.72 + 0.32 = 1.04. Thus, the observation from the plot is correct, that high frustration students receive higher scores. This also suggests a logistic transformation is probably needed.

VarCorr(model\_1)

## user\_id = pdSymm(1)   
## Variance StdDev   
## (Intercept) 0.01057132 0.1028169  
## Residual 0.01337515 0.1156510

The variance from the intercept is 0.01. In other words, a student’s intrapersonal score variance is set to be 0.01.

## Model 2

model\_2 = lme(fixed = month\_avg ~ 1 + mean\_frustratedC,  
 random = list(user\_id = pdSymm(form = ~ 1)),  
 weights = varExp(form = ~ mean\_frustratedC),  
 data = data\_by\_month,  
 na.action = na.exclude,  
 method = 'REML')  
summary(model\_2)

## Linear mixed-effects model fit by REML  
## Data: data\_by\_month   
## AIC BIC logLik  
## -29602.94 -29562.55 14806.47  
##   
## Random effects:  
## Formula: ~1 | user\_id  
## (Intercept) Residual  
## StdDev: 0.1027928 0.1152571  
##   
## Variance function:  
## Structure: Exponential of variance covariate  
## Formula: ~mean\_frustratedC   
## Parameter estimates:  
## expon   
## -2.195122   
## Fixed effects: month\_avg ~ 1 + mean\_frustratedC   
## Value Std.Error DF t-value p-value  
## (Intercept) 0.7186864 0.00216448 21144 332.0363 0  
## mean\_frustratedC 3.1616347 0.26506421 2651 11.9278 0  
## Correlation:   
## (Intr)  
## mean\_frustratedC 0.167   
##   
## Standardized Within-Group Residuals:  
## Min Q1 Med Q3 Max   
## -5.55178368 -0.51440426 0.07521626 0.59177194 4.17989685   
##   
## Number of Observations: 23797  
## Number of Groups: 2653

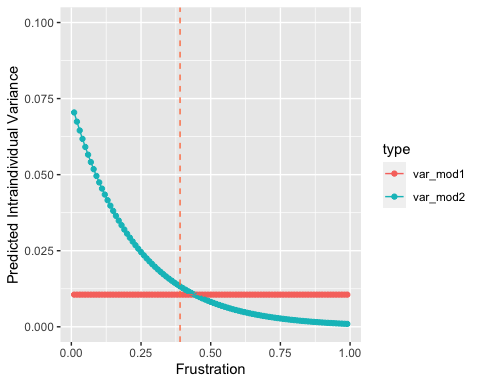
The fixed effects can be interpreted in a similar fashion as above.

Variance is related to frustration negatively. That is, the more frustrated the student is, the less variance we’d predict. So, in general, students who are more frustrated score higher in the assessments in a more consistent way.

Graph the variance in Model 2:

mean\_frustration <- seq(-0.38, 0.60, 0.01) # set up graph x range  
var\_mod2 <- summary(model\_2)$sigma^2 \* exp(coef(2\*model\_2$modelStruct$varStruct, uncons=FALSE)\*mean\_frustration)  
var\_mod1 <- rep(0.01057132, length(mean\_frustration)) # mod 1: variance is constant at 0.01

#make a data frame with predictions and back-transform centered variable to original scale. Both variances in mod 1 and mod 2.  
plotvars <- as.data.frame(cbind(mean\_frustration, var\_mod1, var\_mod2))  
plotvars <- plotvars %>%   
 pivot\_longer(cols = c("var\_mod1", "var\_mod2"), names\_to = "type", values\_to = "value")  
plotvars$mean\_frustration <- plotvars$mean\_frustration + 0.39  
#plotting  
ggplot(data = plotvars, aes(x = mean\_frustration, y = value, color = type), legend=FALSE) +  
 geom\_point() +  
 geom\_line() +   
 xlab("Frustration") +   
 ylab("Predicted Intraindividual Variance") +   
 ylim(0, 0.1) +  
 geom\_vline(xintercept = 0.39, linetype = "dashed", color = "coral")



The graph shows that when the student is in low frustration, the variance of their own scores is estimated higher in model 2 than model 1; conversely, a high frustration student is estimated with a higher intrapersonal variance in model 1 than model 2.

(Note: frustration score only ranges between 0.36 and 0.43 in the data)

# Model comparison

anova(model\_1, model\_2)

## Model df AIC BIC logLik Test L.Ratio p-value  
## model\_1 1 4 -29591.12 -29558.82 14799.56   
## model\_2 2 5 -29602.94 -29562.56 14806.47 1 vs 2 13.81616 2e-04

The ANOVA test shows that allowing heterogeneity in intrapersonal variances significantly increases model fit.

In summary, the analysis finds that students with higher frustration tend to score higher with less variance.