2480 Final Project

Yi Yang

2024-04-11

Package Upload

```
library(readr)
library(haven)
library(psych)
library(tidyverse)
library(labelled)
library(dplyr)
library(haven)
library(tidyverse)
library(ggplot2)
library(lme4)
library(broom)
library(naniar)
library(sjPlot)
library(labelled)
library(performance)
library(knitr)
library(kableExtra)
library(lmerTest)
library(pander)
library(performance)
library(corrplot)
```

Upload Data

```
data <- read_dta("finalproj_2023.dta")
head(data)
```

```
## # A tibble: 6 x 320
##     PID   TAS TAS05 TAS07 TAS09 TAS11 TAS13 TAS15 TAS17 TAS19 ER30000    ER30001
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl+lbl>    <dbl>
## 1  4037     1    NA    NA     1    NA    NA    NA    NA    NA 3 [Releas~       4
## 2  4038     2    NA    NA     1     1    NA    NA    NA    NA 3 [Releas~       4
## 3  4039     5    NA    NA     1     1     1     1     1    NA 3 [Releas~       4
## 4  4041     5    NA    NA    NA     1     1     1     1     1 3 [Releas~       4
## 5  4042     1    NA    NA    NA    NA    NA    NA    NA     1 3 [Releas~       4
## 6  4180     4     1     1     1     1    NA    NA    NA    NA 3 [Releas~       4
## # ... with 308 more variables: ER30002 <dbl>, ER33801 <dbl>, ER33802 <dbl+lbl>,
```

```
## #   ER33803 <dbl+lbl>, ER33804 <dbl>, TA050001 <dbl+lbl>, TA050078 <dbl+lbl>,
## #   TA050676 <dbl+lbl>, TA050679 <dbl+lbl>, TA050686 <dbl+lbl>,
## #   TA050690 <dbl+lbl>, TA050693 <dbl+lbl>, TA050708 <dbl+lbl>,
## #   TA050720 <dbl+lbl>, TA050762 <dbl+lbl>, TA050766 <dbl+lbl>,
## #   TA050770 <dbl+lbl>, TA050778 <dbl+lbl>, TA050786 <dbl+lbl>,
## #   TA050790 <dbl+lbl>, TA050794 <dbl+lbl>, TA050802 <dbl+lbl>, ...
```

Data Cleaning

```
data <- data %>%
 mutate(PID = (ER30001 * 1000) + ER30002) %>%
 relocate(PID) #putting at beginning of dataset
obs <- dim(data)[1]
obs
```

```
## [1] 4776
```

```
sum(duplicated(data$PID))
```

```
## [1] 0
```

```
data$PID <- as.integer(data$PID)
data$Anxiety1<- data$TA050933
data$Anxiety2<- data$TA070914
data$Anxiety3<- data$TA090978
data$Anxiety4<- data$TA111120
data$Anxiety5<- data$TA131212
data$Smoking <- data$TA050762
data$Race <- data$TA050884
data$age1 <- data$ER33804
data$age2 <- data$ER33904
data$age3 <- data$ER34004
data$age4 <- data$ER34104
data$age5 <- data$ER34204


sample_dat <- data %>%
  select(Anxiety1,Anxiety2,Anxiety3,Anxiety4,Anxiety5,
         Smoking, Race, age1,age2,age3,age4,age5,PID) %>%
  dplyr::mutate(Race = case_when(
    Race == 1 ~ "White",
    Race == 2 ~ "Black",
    Race == 3 ~ "Other",
    Race == 4 ~ "Other",
    Race == 5 ~ "Other",
    Race == 7 | Race == 8 | Race == 9 ~ NA_character_
  ))
table(data$Race, useNA = "always")
```

```
##
##    1    2    3    4    5    7    8    9 <NA>
##  378  312    6    8    3    8    2   28 4031
```

3.a Descriptive Statistics of the data

```
head(sample_dat)
```

```
## # A tibble: 6 x 13
##   Anxiety1  Anxiety2 Anxiety3 Anxiety4 Anxiety5 Smoking  Race   age1  age2  age3
##   <dbl+lbl> <dbl+lb> <dbl+lb> <dbl+lb> <dbl+lb> <dbl+lb> <chr> <dbl> <dbl> <dbl>
## 1 NA        NA        1 [Act~ NA       NA       NA       <NA>     18     0    23
## 2 NA        NA        5 [Act~  5 [Act~ NA       NA       <NA>     17     0    22
## 3 NA        NA        3 [Act~  3 [Act~  2 [Act~ NA       <NA>     14    16    18
## 4 NA        NA       NA        3 [Act~  5 [Act~ NA       <NA>     12    14    17
## 5 NA        NA       NA       NA       NA       NA       <NA>      9    10    13
## 6  1 [Actu~  3 [Act~  3 [Act~  3 [Act~ NA        1 [Yes] White    20    22    24
## # ... with 3 more variables: age4 <dbl>, age5 <dbl>, PID <int>
```

```
dim(sample_dat)
```

```
## [1] 4776   13
```

```
des<-sample_dat %>% describe()
# Descriptive statistics of the data
des
```

```
##          vars    n       mean         sd  median    trimmed        mad  min
## Anxiety1    1  745       3.55       1.52       3       3.52       1.48    1
## Anxiety2    2 1115       3.44       1.52       3       3.39       1.48    1
## Anxiety3    3 1554       3.38       1.51       3       3.33       1.48    1
## Anxiety4    4 1907       3.39       1.49       3       3.33       1.48    1
## Anxiety5    5 1804       3.41       1.50       3       3.36       1.48    1
## Smoking     6  745       3.32       2.27       5       3.52       0.00    0
## Race*       7  707       2.09       0.98       3       2.12       0.00    1
## age1        8 4776      11.28       5.53      12      11.44       5.93    0
## age2        9 4776      13.14       5.80      14      13.42       5.93    0
## age3       10 4776      15.14       5.89      16      15.47       5.93    0
## age4       11 4776      16.85       6.32      18      17.34       5.93    0
## age5       12 4776      18.46       6.90      19      19.17       5.93    0
## PID        13 4776 3446254.90 2168046.55 3023504 3439878.81 3126016.14 4037
##              max   range  skew kurtosis       se
## Anxiety1       7       6  0.17    -0.80     0.06
## Anxiety2       7       6  0.30    -0.69     0.05
## Anxiety3       7       6  0.36    -0.54     0.04
## Anxiety4       7       6  0.34    -0.60     0.03
## Anxiety5       7       6  0.30    -0.57     0.04
## Smoking        9       9 -0.57    -1.49     0.08
## Race*          3       2 -0.19    -1.94     0.04
## age1          21      21 -0.22    -0.83     0.08
## age2          23      23 -0.40    -0.46     0.08
## age3          25      25 -0.57     0.04     0.09
## age4          27      27 -0.79     0.59     0.09
## age5          29      29 -1.02     1.06     0.10
## PID      6872174 6868137  0.10    -1.46 31371.58
```

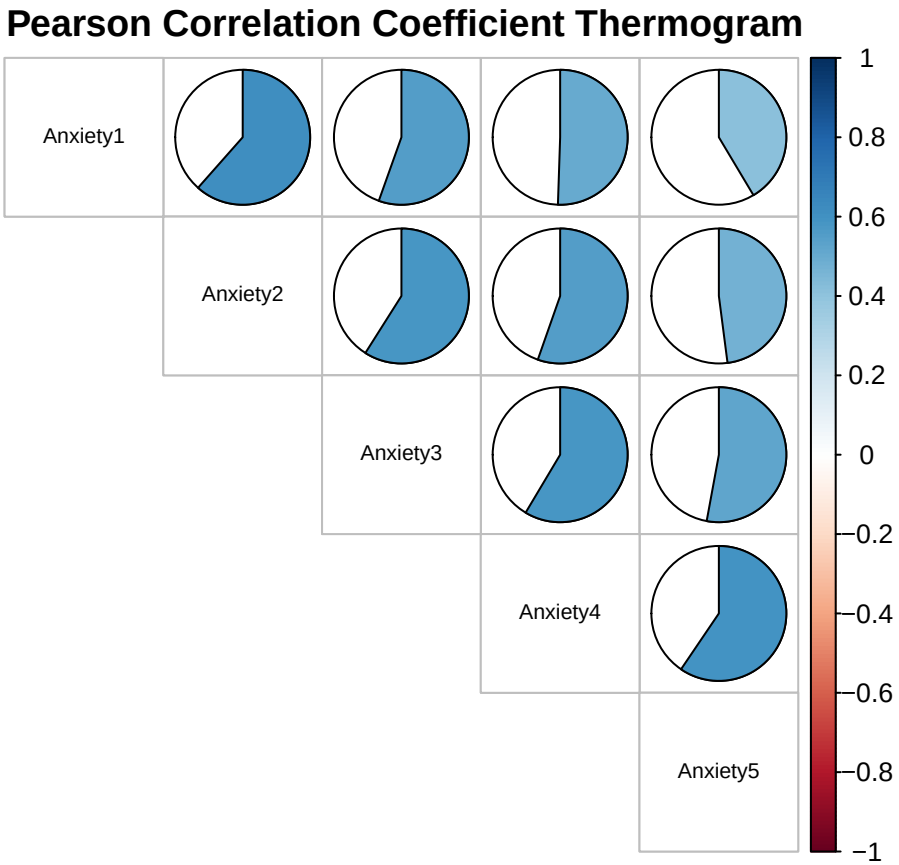
```r
# Calculate the correlation coefficient matrix for Anxiety
cor_anx<-cor(sample_dat[c("Anxiety1","Anxiety2","Anxiety3","Anxiety4", "Anxiety5")],
             use = "pairwise.complete.obs" )
# correlation coefficient matrix for Anxiety
cor_anx
```

```
##           Anxiety1  Anxiety2  Anxiety3  Anxiety4  Anxiety5
## Anxiety1 1.0000000 0.6151934 0.5544389 0.5052706 0.4143576
## Anxiety2 0.6151934 1.0000000 0.5897741 0.5537498 0.4795271
## Anxiety3 0.5544389 0.5897741 1.0000000 0.5855362 0.5289736
## Anxiety4 0.5052706 0.5537498 0.5855362 1.0000000 0.5947919
## Anxiety5 0.4143576 0.4795271 0.5289736 0.5947919 1.0000000
```

```r
corrplot(cor_anx,method='pie',type='upper',tl.col='black',tl.pos='d',tl.cex=0.7,
         show.legend=T, outline=T,insig='p-value',
         title='Pearson Correlation Coefficient Thermogram', mar=c(0,0,1,0))
```

```
## Warning in text.default(pos.ylabel[, 1] + 0.5, pos.ylabel[, 2],
## newcolnames[1:min(n, : "show.legend" is not a graphical parameter
```

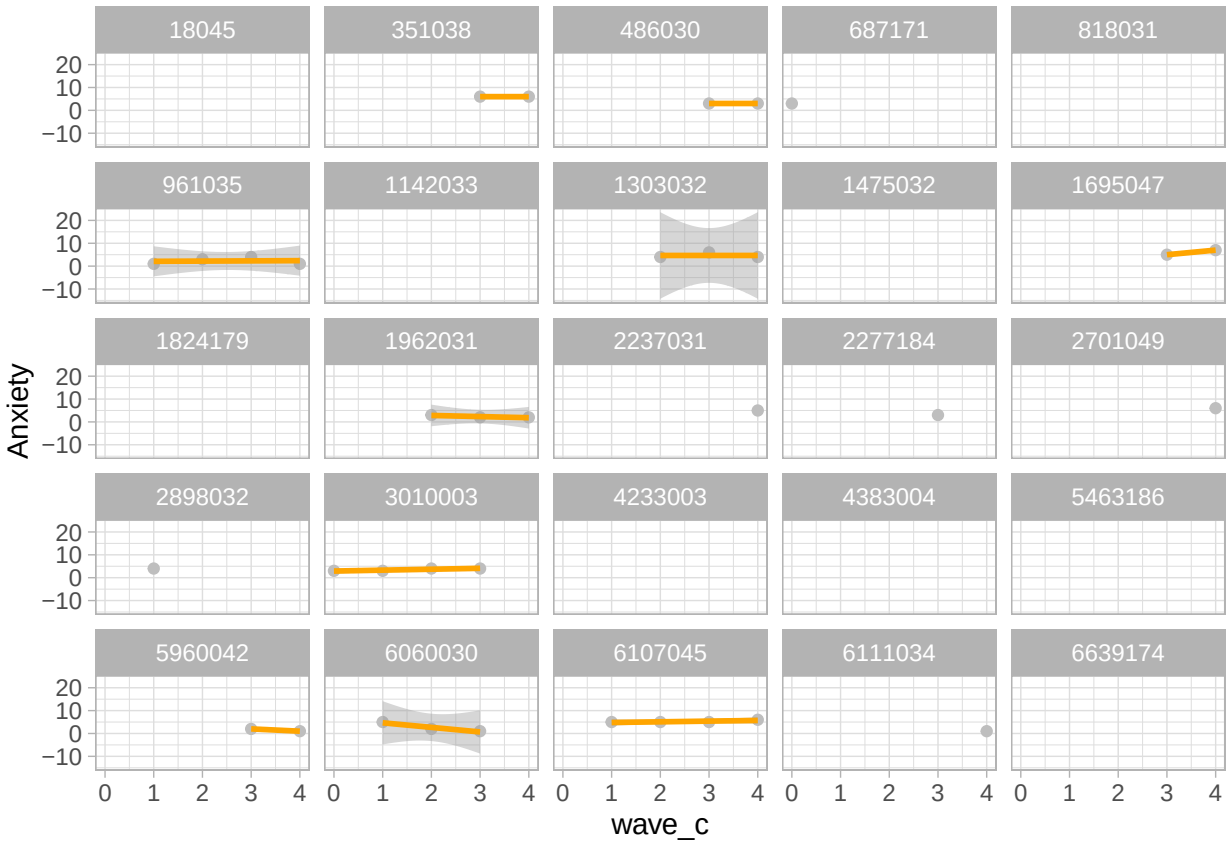
```
## Warning in title(title, ...): "show.legend" is not a graphical parameter
```



3.b i Describe the growth in your outcome

```
obs <- dim(sample_dat)[1] # data size
set.seed(0)
sample_data <- sample_dat[sample(obs, size = 25),] # Sampling of 25 samples
sample_dat_long <- sample_data %>%
  select(Anxiety1,Anxiety2,Anxiety3,Anxiety4,Anxiety5,PID,Race) %>%
  pivot_longer(cols = c("Anxiety1","Anxiety2","Anxiety3","Anxiety4","Anxiety5"),
               values_to = "Anxiety") %>% mutate(wave = case_when(
                 name == "Anxiety1" ~ 1,
                 name == "Anxiety2" ~ 2,
                 name == "Anxiety3" ~ 3,
                 name == "Anxiety4" ~ 4,
                 name == "Anxiety5" ~ 5))
sample_dat_long$wave_c <- (sample_dat_long$wave) - 1
#Individual growth plots
ggplot(data = sample_dat_long, aes(x = wave_c, y = Anxiety)) +
  geom_point(col='gray') + geom_smooth(method = "lm",col='orange') +
  facet_wrap(vars(PID))+theme_light()
```

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



3.b ii Individual OLS regressions conducted and visualized with the mean trajectory line.

```r
#Individual parametric trajectories with mean OLS trajectory
ggplot(data = sample_dat_long, aes(x = wave_c, y = Anxiety)) +
  geom_smooth(aes(group = as.factor(PID)), method = "lm", color="gray",cex=0.8,se=F) +
  geom_smooth(method = "lm",color ="orange",se=F,cex=0.9,lty=6)+
  labs(x="Wave",y="Anxiety",title="Individual parametric trajectories with mean OLS trajectory")+
  theme_light()
```

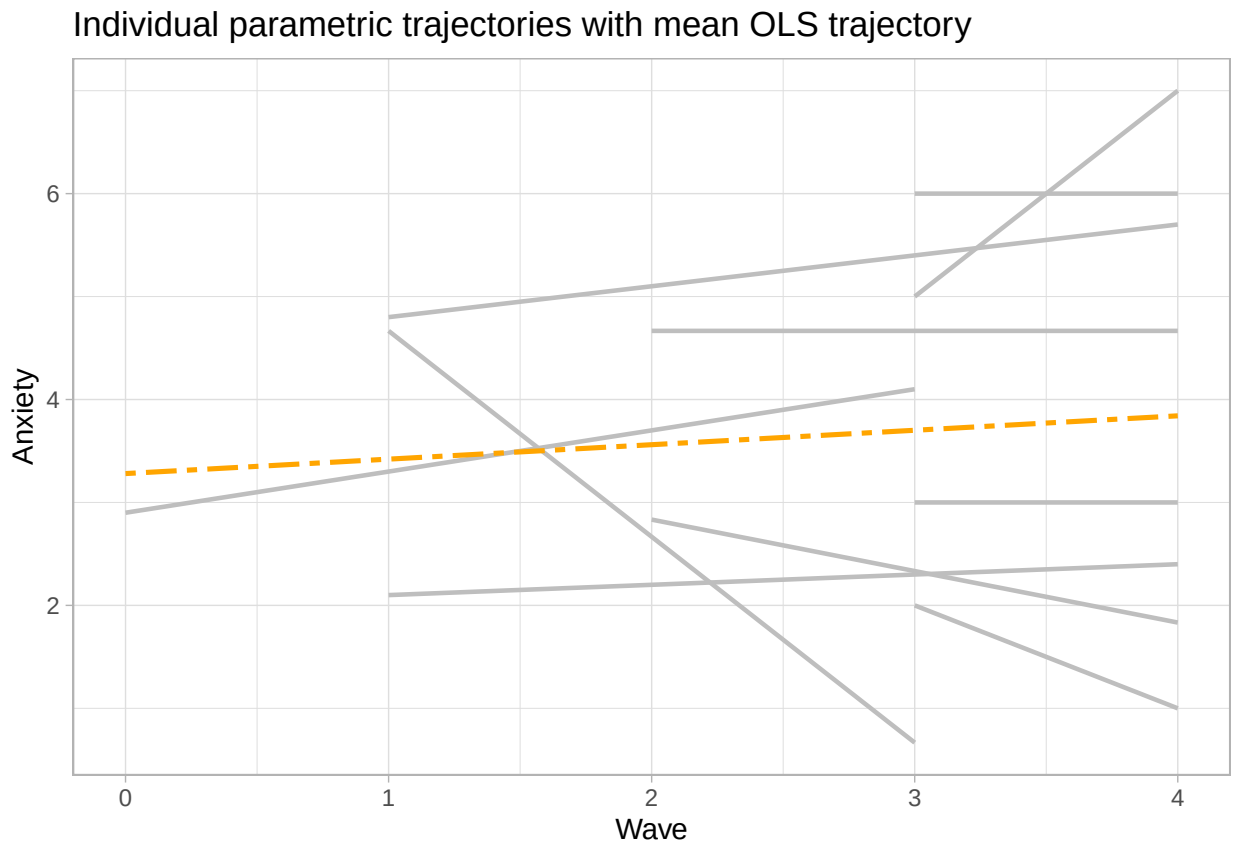
```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.

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

## Warning: Removed 90 rows containing non-finite values ('stat_smooth()').

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

## Warning: Removed 90 rows containing non-finite values ('stat_smooth()').
```



3.c i sample means of the estimated intercepts and slopes

```r
sample_dat_long <- sample_dat %>%
  select(Anxiety1,Anxiety2,Anxiety3,Anxiety4,Anxiety5,PID) %>%
  pivot_longer(cols = c("Anxiety1","Anxiety2","Anxiety3","Anxiety4","Anxiety5"),
               values_to = "Anxiety") %>% mutate(wave = case_when(
                 name == "Anxiety1" ~ 1,
                 name == "Anxiety2" ~ 2,
                 name == "Anxiety3" ~ 3,
                 name == "Anxiety4" ~ 4,
                 name == "Anxiety5" ~ 5))
sample_dat_long$wave_c <- sample_dat_long$wave - 1

# Group by PID and create a new missing wave column
sample_dat_long_2 <- sample_dat_long %>%
  group_by(PID) %>%
  dplyr::mutate(missing_waves = sum(is.na(Anxiety)))

# Group by PID and filter for missing_wave less than 3
sample_dat_long3 <- sample_dat_long_2 %>%
  group_by(PID) %>%
  filter(sum(missing_waves) < 3)

# Building a linear model
model1 <- sample_dat_long3  %>% dplyr::group_by(PID) %>%
  do(model = lm(Anxiety ~ wave_c, data =.))
model1[[2]][[1]]
```

```
##
## Call:
## lm(formula = Anxiety ~ wave_c, data = .)
##
## Coefficients:
## (Intercept)       wave_c
##         2.2          0.8
```

```r
intercept <- slope <- NULL

# Calling slope and intercept
for(i in 1:nrow(model1)){
  intercept[i] <- model1[[2]][[i]][["coefficients"]][1]
  slope[i] <- model1[[2]][[i]][["coefficients"]][2]
}
```

3.c ii Sample Variance

```r
summary(intercept)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   2.600   3.400   3.575   4.400   6.800
```

```r
var(intercept)
```

```
## [1] 1.932864
```

```r
summary(slope)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1.20000 -0.30000 -0.10000 -0.07017  0.17500  1.10000
```

```r
var(slope)
```

```
## [1] 0.1469122
```

3.c iii correlation between the estimated intercepts and slopes

```r
# Check the covariance of slope and intercept
cor(intercept,slope)
```

```
## [1] -0.5745468
```

3.D Model building

3.D.i Conduct the unconditional mean model

```r
dat_long <- sample_dat %>%
  dplyr::select(Anxiety1,Anxiety2,Anxiety3,Anxiety4,Anxiety5,
          age1,age2,age3,age4,age5,PID)  %>%
  pivot_longer(-PID) %>%
  separate(name, into = c("name", "wave"), sep = "(?<=[A-Za-z])(?=[0-9])") %>%
  pivot_wider(names_from = "name", values_from = "value")

dat_mari_race <- sample_dat %>%
  select(Smoking, Race,PID)

dat_long <- left_join(dat_long, dat_mari_race, by = "PID")

dat_long <- remove_labels(dat_long)

dat_long$wave_c <- as.integer(dat_long$wave)-1
table(dat_long$Race)
```

3.D.i 1 Interpret the fixed and random effects

```
##
## Black Other White
##  1560    85  1890
```

```
model.a <- lmer(Anxiety ~ 1 + (1 |PID), data = dat_long, REML = FALSE)
summary(model.a)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: Anxiety ~ 1 + (1 | PID)
##    Data: dat_long
##
##      AIC      BIC   logLik deviance df.resid
##  23929.8  23950.4 -11961.9  23923.8     7122
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1412 -0.5761 -0.0707  0.5550  3.9357
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  PID      (Intercept) 1.274    1.129
##  Residual             1.003    1.001
## Number of obs: 7125, groups:  PID, 2570
##
## Fixed effects:
##              Estimate Std. Error        df t value Pr(>|t|)
## (Intercept) 3.431e+00  2.587e-02 2.508e+03   132.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
performance::icc(model.a)
```

3.D.i 2Conduct the ICC and interpret

```
## # Intraclass Correlation Coefficient
##
##     Adjusted ICC: 0.559
##   Unadjusted ICC: 0.559
```

```
icc_n <- as.data.frame(VarCorr(model.a),comp="Variance")$vcov[1]
icc_d <- as.data.frame(VarCorr(model.a),comp="Variance")$vcov[1] +
         as.data.frame(VarCorr(model.a),comp="Variance")$vcov[2]
icc_n / icc_d
```

```
## [1] 0.5594694
```

3.D.ii Conduct the unconditional growth model

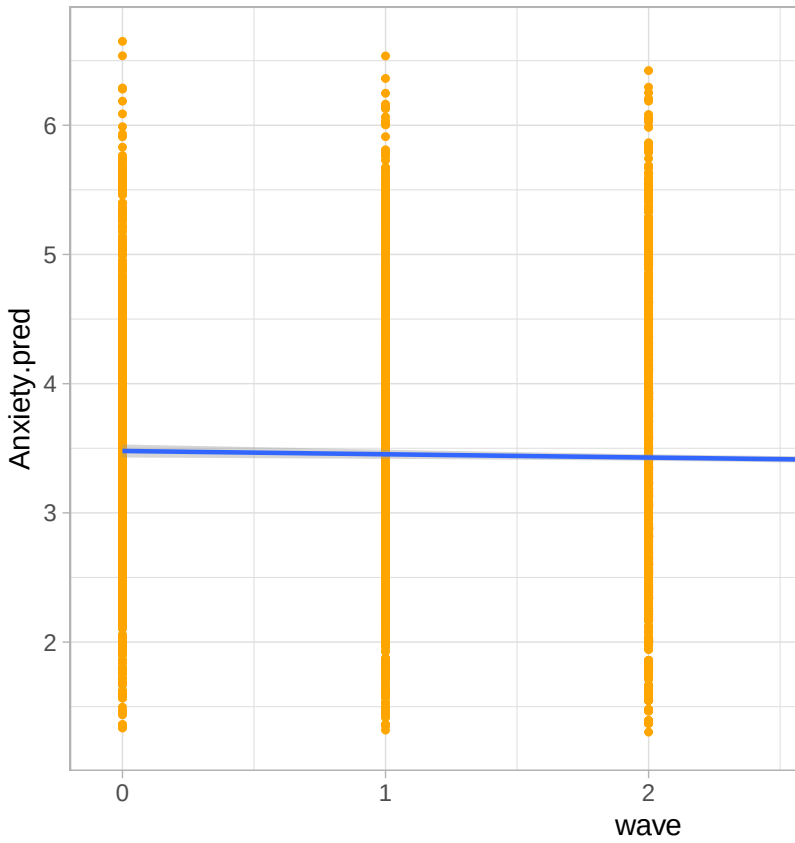
```
model.b <- lmer(Anxiety ~ wave_c + (wave_c|PID), data = dat_long, REML = FALSE)
summary(model.b)
```

3.D.ii 1 Interpret the fixed and random effects

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: Anxiety ~ wave_c + (wave_c | PID)
##    Data: dat_long
##
##      AIC      BIC   logLik deviance df.resid
##  23865.5  23906.8 -11926.8  23853.5     7119
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.3626 -0.5509 -0.0683  0.5270  4.1238
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  PID      (Intercept) 1.56071  1.2493
##           wave_c      0.05378  0.2319   -0.42
##  Residual             0.90482  0.9512
## Number of obs: 7125, groups:  PID, 2570
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)    3.55926    0.04012 1545.91034  88.717   <2e-16 ***
## wave_c        -0.05030    0.01176 1591.66268  -4.278    2e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##        (Intr)
## wave_c -0.763
## optimizer (nloptwrap) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.00603926 (tol = 0.002, component 1)
```

```
data_tmp <- data.frame(Anxiety.pred = predict(model.b),
                       wave = model.b@frame[["wave_c"]])

ggplot(data = data_tmp, mapping = aes(x = wave, y = Anxiety.pred)) +
  geom_point(col='orange',cex=0.9) +
  stat_smooth(method="lm", formula = y ~ x,cex=0.8) +
  theme_light()
```



3.D.ii 2 Graph the unconditional growth model