

project5

Ryan Wei

2024-05-02

```
dat <- read_excel("./Proj 4 data.xlsx")

dat$school <- as.factor(dat$school)
dat$group <- as.factor(dat$group)
dat$ID <- as.factor(dat$ID)
dat <- dat %>%
  group_by(ID) %>%
  arrange(time) %>%
  mutate(SFD_baseline = first(SFD), observed = factor(as.numeric(!is.na(SFD))), n_miss = sum(is.na(SFD)))
  ungroup %>%
  mutate(time = factor(time))
```

```
library(table1)
```

```
##
## Attaching package: 'table1'

## The following objects are masked from 'package:base':
##
##      units, units<-
```

```
table1(~ SFD_baseline + school | group, data=dat %>% filter(time == 1))
```

	0	1	Overall
	(N=49)	(N=59)	(N=108)
SFD_baseline			
Mean (SD)	9.60 (4.64)	8.95 (4.77)	9.25 (4.70)
Median [Min, Max]	11.0 [0, 14.0]	11.0 [0, 14.0]	11.0 [0, 14.0]
Missing	1 (2.0%)	2 (3.4%)	3 (2.8%)
school			
1	0 (0%)	28 (47.5%)	28 (25.9%)
2	21 (42.9%)	0 (0%)	21 (19.4%)
3	0 (0%)	31 (52.5%)	31 (28.7%)
4	28 (57.1%)	0 (0%)	28 (25.9%)

```
traj.plot<-
  dat %>%
  ggplot(aes(x = time, y = SFD)) +
```

```

geom_line(alpha = 0.3, linewidth = 0.5, aes(group = ID, color = group)) +
facet_grid(.~school) +
geom_smooth(color="blue", method="lm", aes(group = 1), se = FALSE) +
ylab("SFD (in last two weeks)") +
xlab("Follow up time (months)") +
scale_x_discrete(labels = c(0,6,12)) +
labs(color = "Treatment Group") +
theme_bw() +
theme(legend.position = "none")

traj.box <-
dat %>%
ggplot(aes(x = time, y = SFD, color = group)) +
geom_boxplot(alpha = 0.3, linewidth = 0.5) +
facet_grid(.~school) +
ylab("SFD (in last two weeks)") +
xlab("Follow up time (months)") +
scale_x_discrete(labels = c(0,6,12)) +
labs(color = "Treatment Group") +
theme_bw() +
theme(legend.position = "none")

ggsave("./plots/trajectory_plot.jpg", (traj.plot/traj.box) + theme(legend.position = "bottom"), width =

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

## Warning: Removed 28 rows containing non-finite outside the scale range
## ('stat_smooth()').

## Warning: Removed 26 rows containing missing values or values outside the scale range
## ('geom_line()').

## Warning: Removed 28 rows containing non-finite outside the scale range
## ('stat_boxplot()').

traj.miss <-
dat %>% group_by(time, group) %>%
mutate(sum_miss = sum(observed == 0), total_obs = n(), prop_miss = sum_miss/ total_obs) %>%
ggplot(aes(x = time, y = prop_miss, fill = group)) +
facet_grid(. ~ school) +
geom_bar(stat = "identity", position = position_dodge()) +
labs(y = "Proportion of Missing Observations", x = "Group", color = "Treatment Group") +
theme_bw()

ggsave("./plots/missing_plot.jpg", traj.miss, width = 10, height = 8, dpi = 300)

mean.score.plot <-
dat %>%
group_by(ID) %>%
mutate(total_obs = sum(as.numeric(observed))) %>%
ungroup() %>%

```

```

mutate(type = as.factor(ifelse(total_obs == 3, "Completer", "Drop-out"))) %>%
group_by(type, time, group) %>%
mutate(mean_SFD = mean(SFD, na.rm = T)) %>%
dplyr::select(time, group, type, mean_SFD, school) %>%
distinct() %>%
ungroup() %>%
ggplot(aes(x = time, y = mean_SFD, group = interaction(group, type), color = group, linetype = type))
geom_line(alpha = 0.8, size = 0.5) +
scale_linetype_manual(values = c("Completer" = "solid", "Drop-out" = "dashed")) +
labs(color = "Treatment Group", linetype = "Type") +
ylab("SFD") +
xlab("Follow up time (months)") +
scale_x_discrete(labels = c(0, 6, 12)) +
theme_bw()

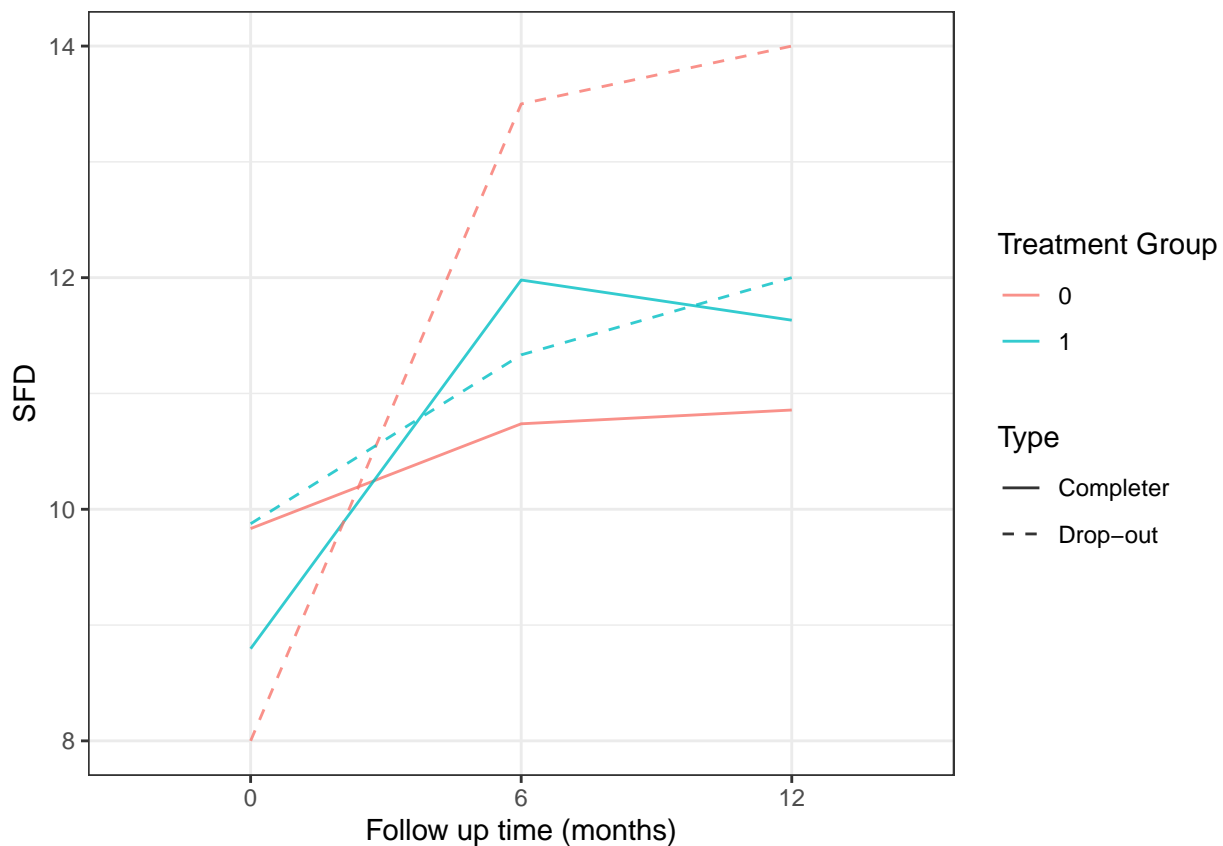
```

```

## 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.

```

```
mean.score.plot
```



```
ggsave("./plots/mean_SFD_plot.jpg", mean.score.plot, width = 12, height = 8, dpi = 300)
```

Zero-inflated Negative Binomial model

excluding subjects with missing at all time points

```
dat.comp <- dat %>% filter(n_miss < 3, time != 1) %>%  
  mutate(time = factor(time, levels = c(2,3)))  
mdl.nb <- glmer.nb(SFD ~ group*time + SFD_baseline +  
  (1 | school), data = dat.comp, control = glmerControl(optimizer = "bobyqa"))
```

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

```
dat %>% filter(time == "1") %>% group_by(school) %>%  
  summarise(mean_SFD = mean(SFD, na.rm = T), var_SFD = var(SFD, na.rm = T))
```

```
## # A tibble: 4 x 3  
##   school mean_SFD var_SFD  
##   <fct>     <dbl>   <dbl>  
## 1 1      9.96    18.8  
## 2 2      9.38    25.7  
## 3 3      8.03    25.2  
## 4 4      9.78    19.1
```

```
dat %>% filter(time == "2") %>% group_by(school) %>%  
  summarise(mean_SFD = mean(SFD, na.rm = T), var_SFD = var(SFD, na.rm = T))
```

```
## # A tibble: 4 x 3  
##   school mean_SFD var_SFD  
##   <fct>     <dbl>   <dbl>  
## 1 1     11.7    15.8  
## 2 2     11.8     6.40  
## 3 3     12.1     3.67  
## 4 4     10.4    31.0
```

```
dat %>% filter(time == "3") %>% group_by(school) %>%  
  summarise(mean_SFD = mean(SFD, na.rm = T), var_SFD = var(SFD, na.rm = T))
```

```
## # A tibble: 4 x 3  
##   school mean_SFD var_SFD  
##   <fct>     <dbl>   <dbl>  
## 1 1     11.7    11.9  
## 2 2     11.0    18.7  
## 3 3     11.6    14.7  
## 4 4     10.9    21.4
```

```
summary(mdl.nb)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: Negative Binomial(17.8746) ( log )
## Formula: SFD ~ group * time + SFD_baseline + (1 | school)
## Data: dat.comp
## Control: glmerControl(optimizer = "bobyqa")
##
##      AIC      BIC    logLik deviance df.resid
##  1180.7   1203.5   -583.4   1166.7     184
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.7156 -0.2054  0.2746  0.5366  1.2650
##
## Random effects:
## Groups Name      Variance Std.Dev.
## school (Intercept) 0          0
## Number of obs: 191, groups: school, 4
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.237336   0.080645  27.743 < 2e-16 ***
## group1       0.095933   0.076870   1.248  0.21204
## time3       -0.011868   0.081491  -0.146  0.88421
## SFD_baseline 0.016396   0.005914   2.773  0.00556 **
## group1:time3 -0.016639   0.110307  -0.151  0.88010
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##              (Intr) group1 time3  SFD_bs
## group1      -0.556
## time3       -0.464  0.509
## SFD_baselin -0.713  0.056 -0.033
## group1:tim3  0.349 -0.694 -0.738  0.015
## optimizer (bobyqa) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

```
mdl.zinb.1 <- glmmTMB(SFD ~ group*time + SFD_baseline + (1|school) ,
                     ziformula = ~ 1 ,
                     data = dat.comp,
                     family = nbinom2,
                     control=glmmTMBControl(optimizer=optim,optArgs=list(method="BFGS")))

summary(mdl.zinb.1)
```

```
## Family: nbinom2 ( log )
## Formula:          SFD ~ group * time + SFD_baseline + (1 | school)
## Zero inflation:    ~1
## Data: dat.comp
##
##      AIC      BIC    logLik deviance df.resid
##   958.2   984.2   -471.1    942.2     183
##
```

```
## Random effects:
##
## Conditional model:
##   Groups Name      Variance Std.Dev.
##   school (Intercept) 4.239e-07 0.000651
## Number of obs: 191, groups:  school, 4
##
## Dispersion parameter for nbinom2 family (:): 6.56e+04
##
## Conditional model:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.463080   0.063927  38.53   <2e-16 ***
## group1       0.011756   0.060020   0.20   0.845
## time3       0.001518   0.064152   0.02   0.981
## SFD_baseline 0.004942   0.004712   1.05   0.294
## group1:time3 -0.006779   0.086247  -0.08   0.937
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Zero-inflation model:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.4630   0.2691  -9.154   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mdl.zinb.2 <- glmmTMB(SFD ~ group*time + SFD_baseline + (1|school) ,
                     zi = ~ group,
                     data = dat.comp,
                     family = nbinom2,
                     control=glmmTMBControl(optimizer=optim,optArgs=list(method="BFGS")))

summary(mdl.zinb.2)
```

```
## Family: nbinom2 ( log )
## Formula:          SFD ~ group * time + SFD_baseline + (1 | school)
## Zero inflation:    ~group
## Data: dat.comp
##
##      AIC      BIC   logLik deviance df.resid
##    957.6    986.8   -469.8    939.6     182
##
## Random effects:
##
## Conditional model:
##   Groups Name      Variance Std.Dev.
##   school (Intercept) 4.396e-08 0.0002097
## Number of obs: 191, groups:  school, 4
##
## Dispersion parameter for nbinom2 family (:): 1.85e+05
##
## Conditional model:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.463145   0.063921  38.53   <2e-16 ***
## group1       0.011738   0.060012   0.20   0.845
```

```
## time3          0.001460    0.064148    0.02    0.982
## SFD_baseline  0.004938    0.004712    1.05    0.295
## group1:time3 -0.006714    0.086241   -0.08    0.938
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Zero-inflation model:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.0669     0.3357  -6.158 7.38e-10 ***
## group1      -0.8987     0.5684  -1.581  0.114
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova mdl.zinb.1, mdl.zinb.2)
```

```
## Data: dat.comp
## Models:
## mdl.zinb.1: SFD ~ group * time + SFD_baseline + (1 | school), zi=~1, disp=~1
## mdl.zinb.2: SFD ~ group * time + SFD_baseline + (1 | school), zi=~group, disp=~1
##           Df      AIC      BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## mdl.zinb.1  8 958.23 984.25 -471.11  942.23
## mdl.zinb.2  9 957.56 986.83 -469.78  939.56 2.6721      1      0.1021
```

```
# null model
mdl.zinb.null <- glmmTMB(SFD ~ time + SFD_baseline + (1|school) ,
                        zi = ~ group,
                        data = dat.comp,
                        family = nbinom2,
                        control=glmmTMBControl(optimizer=optim,optArgs=list(method="BFGS")))

anova(mdl.zinb.2, mdl.zinb.null)
```

```
## Data: dat.comp
## Models:
## mdl.zinb.null: SFD ~ time + SFD_baseline + (1 | school), zi=~group, disp=~1
## mdl.zinb.2: SFD ~ group * time + SFD_baseline + (1 | school), zi=~group, disp=~1
##           Df      AIC      BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## mdl.zinb.null  7 953.60 976.36 -469.80  939.60
## mdl.zinb.2     9 957.56 986.83 -469.78  939.56 0.0428      2      0.9788
```

```
# random effects in zi model
mdl.zinb.3 <- glmmTMB(SFD ~ time + SFD_baseline + (1|school) ,
                    zi = ~ group + (1|school),
                    data = dat.comp,
                    family = nbinom2,
                    control=glmmTMBControl(optimizer=optim,optArgs=list(method="BFGS")))

anova(mdl.zinb.2, mdl.zinb.3)
```

```
## Data: dat.comp
## Models:
## mdl.zinb.3: SFD ~ time + SFD_baseline + (1 | school), zi=~group + (1 | school), disp=~1
```

```
## mdl.zinb.2: SFD ~ group * time + SFD_baseline + (1 | school), zi=~group, disp=~1
##          Df      AIC      BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
## mdl.zinb.3  8 955.60 981.62 -469.80   939.60
## mdl.zinb.2  9 957.56 986.83 -469.78   939.56 0.0431      1    0.8356
```