P8131_hw7

Shan Jiang

Import data

```
library(tidyverse)
library(readxl)
library(lme4)
library(nlme)
health  = readxl::read_xlsx("./HW8-HEALTH.xlsx")
head(health)
```

```
## # A tibble: 6 x 5
##      ID  TIME TXT          HEALTH AGEGROUP
##   <dbl> <dbl> <chr>        <chr>  <chr>
## 1   101     1 Intervention Good   15-24
## 2   101     2 Intervention Good   15-24
## 3   101     3 Intervention Good   15-24
## 4   101     4 Intervention Good   15-24
## 5   102     1 Control      Poor   15-24
## 6   102     2 Control      Poor   15-24
```

```
## factorize the variable HEALTH, TXT
health$HEALTH = as.factor(health$HEALTH)
levels(health$HEALTH)
```

```
## [1] "Good" "Poor"
```

```
health$TXT = as.factor(health$TXT)
levels(health$TXT)
```

```
## [1] "Control"      "Intervention"
```

```
##  n=80:  4 points in time: randomization, 3 months, 6 months, and 12 months post-randomization.
```

(a) Baseline comparison

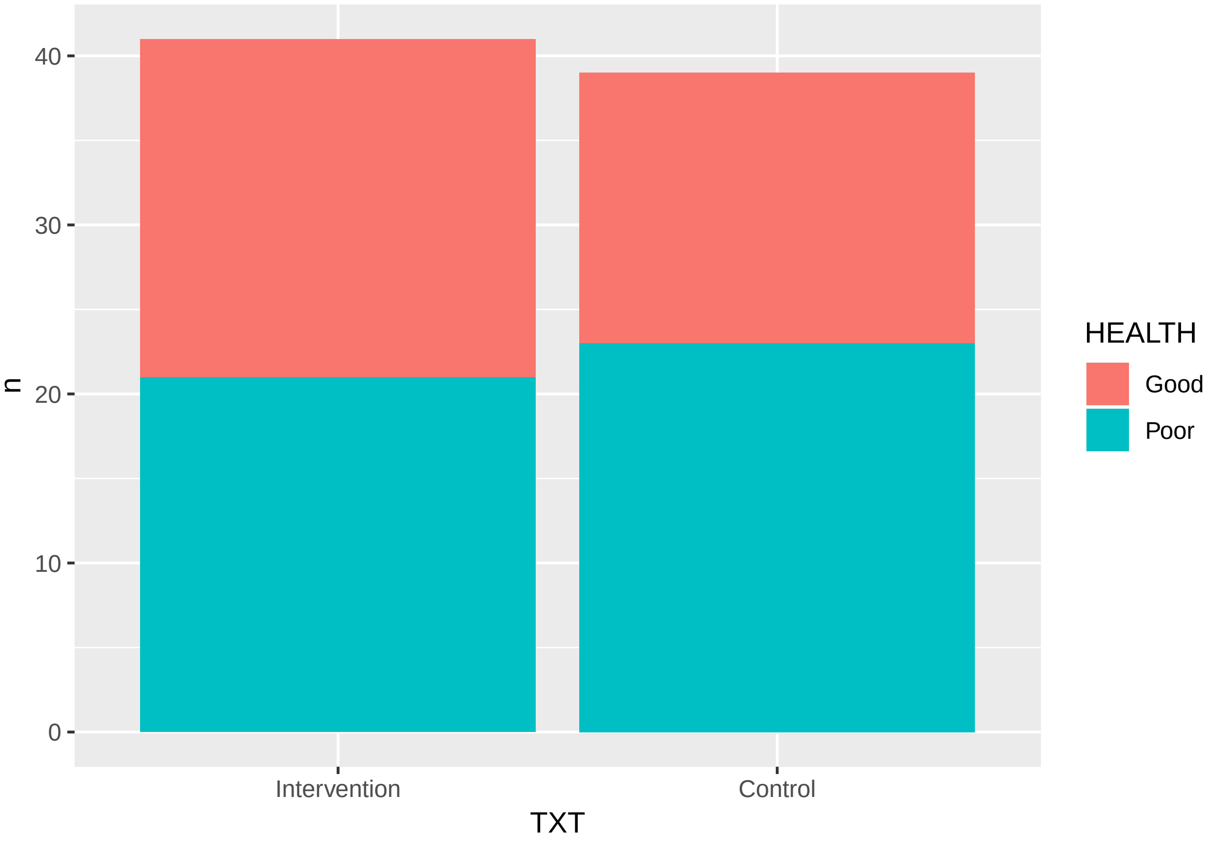
Boxplots to show the relationship

1. Descriptive statistics

```
## subset data which only have value at the baseline level
health.sub <- subset(health, TIME == "1")
nrow(health.sub)
```

```
## [1] 80
```

```
## There are in total 80 baseline individual rows in our dataset
## Barplot
health.sub %>%
  group_by(TXT, HEALTH) %>%
  summarize(n = n()) %>%
  ggplot(., aes(x = TXT, y = n, fill = HEALTH)) +
  geom_bar(stat = "identity") +
  scale_x_discrete(labels = c("Intervention","Control"))
```



From this stacked barplot, we can see a slight difference in the baseline status between the intervention group and control group, with control group has a higher poor health status subjects in their pool while the intervention group subjects are more balanced in their health status.

2. Quantatitive comparison

```
# Fit a simple GLM model
hea_glm <- glm(HEALTH ~ TXT, data = health.sub , family = "binomial")
summary(hea_glm)
```

```
##
## Call:
## glm(formula = HEALTH ~ TXT, family = "binomial", data = health.sub)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.335  -1.198   1.028   1.157   1.157
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.04879    0.31244   0.156    0.876
## TXTIntervention  0.31412    0.45122   0.696    0.486
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 110.10  on 79  degrees of freedom
## Residual deviance: 109.62  on 78  degrees of freedom
## AIC: 113.62
##
## Number of Fisher Scoring iterations: 4
```

The GLM model shows that the p-value is 0.486, which is greater than 0.05, so there is no significant difference between the control group and intervention group in their subjects' health status.

(b) GEE model

```r
library(gee) # Note: data need to be sorted!!! make sure measures from the same subject are together!
dim(health) # there are 279 datapoints, and it is an unbalanced design as there are some data points missing
```

```
## [1] 279   5
```

reconstruct the variable: baseline status and month

```r
# make time 1 as another covariate: baseline status
health = health %>%
  group_by(ID) %>%
  mutate(HEALTH = as.numeric(HEALTH == "Good") ) %>%
   mutate(baseline = ifelse(TIME == "1", HEALTH[TIME =="1"], HEALTH[TIME =="1"])) %>%
   mutate(treatment = TXT) %>%
   select(-TXT)

health$AGEGROUP = as.factor(health$AGEGROUP)
levels(health$AGEGROUP)
```

```
## [1] "15-24" "25-34" "35+"
```

```r
health$baseline = recode(health$baseline, "1" = "0", "0" = "1")


### month post randomization: months
health = health %>%
  mutate(month = if_else(TIME == "4", (TIME - 1) * 4, (TIME - 1) * 3))
```

Fit model GEE.

```r
# Unstructured
health.follow <- subset(health, month > "0")

hea_gee1 <- gee(HEALTH ~ month + treatment + baseline + AGEGROUP, data = health.follow,
                 family = "binomial", id = ID,
                 corstr = "unstructured",
                 scale.fix = TRUE, scale.value = 1) # scale parameter is phi (over dispersion)
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

```
## running glm to get initial regression estimate
```

```
##           (Intercept)                 month treatmentIntervention
##            0.18528086            0.02536275            1.99669985
##             baseline1         AGEGROUP25-34           AGEGROUP35+
##           -1.71063852            1.19749448            1.39742621
```

```r
summary(hea_gee1)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Logit
##  Variance to Mean Relation: Binomial
##  Correlation Structure:     Unstructured
##
## Call:
## gee(formula = HEALTH ~ month + treatment + baseline + AGEGROUP,
##     id = ID, data = health.follow, family = "binomial", corstr = "unstructured",
##     scale.fix = TRUE, scale.value = 1)
##
## Summary of Residuals:
##         Min          1Q      Median          3Q         Max
## -0.98144969 -0.18317233  0.08914345  0.17159228  0.83093959
##
##
## Coefficients:
##                          Estimate Naive S.E.    Naive z Robust S.E.
## (Intercept)            0.12457924 0.47137316  0.2642901  0.51374172
## month                  0.03243343 0.03665686  0.8847848  0.04755408
## treatmentIntervention  2.10225898 0.48779381  4.3097286  0.53777951
## baseline1             -1.81418056 0.48958528 -3.7055456  0.50961334
## AGEGROUP25-34          1.35250468 0.48130172  2.8100973  0.50420159
## AGEGROUP35+            1.42052166 0.79781620  1.7805124  0.78372968
##                         Robust z
## (Intercept)            0.2424939
## month                  0.6820326
## treatmentIntervention  3.9091467
## baseline1             -3.5599158
## AGEGROUP25-34          2.6824681
## AGEGROUP35+            1.8125148
##
## Estimated Scale Parameter:  1
## Number of Iterations:  5
##
## Working Correlation
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.1719328 0.5859907
## [2,] 0.1719328 1.0000000 0.2013998
## [3,] 0.5859907 0.2013998 1.0000000
```

Exponentiate values back.

```r
## values reflect effects on the log-odds scale
exp(coef(hea_gee1)["month"])
```

```
##    month
## 1.032965
```

```r
exp(coef(hea_gee1)["treatmentIntervention"])
```

```
## treatmentIntervention
##              8.184638
```

```
exp(coef(hea_gee1)["baseline"])
```

```
## <NA>
##   NA
```

```
exp(coef(hea_gee1)["AGEGROUP25-34"])
```

```
## AGEGROUP25-34
##      3.867099
```

```
exp(coef(hea_gee1)["AGEGROUP35+"])
```

```
## AGEGROUP35+
##    4.139279
```

- GEE model:
- Interpret your results:
- Intercept: the log odds of good health rating is 0.12457924 on average for patients with Good baseline health rating and age within 15-24 at control group at baseline.
- $\beta Month$: For patients within the same age group and same treatment group and same baseline rating, one unit increase of month is associated with 0.03243 increase in log odds ratio in being at good health status when keeping all variables constant.
- $\beta_{Age25-34}$: The expected value of log odds ratio is 1.3525 being at good health status in the AGE GROUP of 25-34 vs these who do not fall into the age group, adjusted for all other covariates for patients within the same treatment group and same baseline rating at same post randomization month.
- $\beta_{AGEGROUP35+}$: : For patients within the same treatment group and same baseline rating at same post randomization month, the log odds ratio of good health rating for age above 35 vs. age within 15-24 is 1.42052166 on average, adjusted for all other covariates.
- $\beta_{Treatment}$: The expected value of log odds ratio for being at good health status is 2.10226 for these who have intervention vs these who are of control, adjusted for all other covariates.
- $\beta_{baseline1}$: The expected value of log odds ratio is -1.8141 of being at good health status for subject who are of good health status at the baseline compared with these who are of poor health status at the baseline, adjusted for all other covariates.

(c) Fit a GLMM.

```
# fit GLMM
# random intercept model

h.GLMM1 <- glmer(HEALTH  ~ baseline + AGEGROUP +  treatment + month + (1 | ID),
                 family = 'binomial', data =  health.follow)

summary(h.GLMM1)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: HEALTH ~ baseline + AGEGROUP + treatment + month + (1 | ID)
##    Data: health.follow
##
##      AIC      BIC   logLik deviance df.resid
##    185.0    208.0    -85.5    171.0      192
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.6112 -0.2327  0.1402  0.2982  1.8239
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  ID     (Intercept) 5.721    2.392
## Number of obs: 199, groups:  ID, 78
##
## Fixed effects:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            0.19521    0.87019   0.224  0.82250
## baseline1             -2.77610    0.98381  -2.822  0.00478 **
## AGEGROUP25-34          2.25651    1.00877   2.237  0.02529 *
## AGEGROUP35+            1.98229    1.38118   1.435  0.15123
## treatmentIntervention  3.41325    1.07268   3.182  0.00146 **
## month                  0.03718    0.06933   0.536  0.59176
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) basln1 AGEGROUP2 AGEGROUP3 trtmnI
## baseline1   -0.374
## AGEGROUP25- -0.319 -0.379
## AGEGROUP35+ -0.195 -0.274  0.390
## trtmntIntrv -0.256 -0.449  0.395     0.206
## month       -0.472 -0.016  0.007    -0.007     0.047
```

```
# correlation of fixed effects is related to Fisher information of estimates
random.effects(h.GLMM1)
```

```
## $ID
##     (Intercept)
## 101  0.27142497
## 102 -0.76011560
## 103  0.60548770
## 104  0.03645967
## 105 -0.31268468
## 106  2.06236161
## 107  1.51701009
## 109  0.03645967
## 110  1.73646579
## 111  0.03645967
## 112 -2.31907987
## 113 -0.55820367
## 114  0.39872805
## 116  0.48079400
## 117 -2.42575706
## 118 -0.91824620
## 119  0.31381047
## 120 -1.56581860
## 121 -0.55820367
## 122  0.39872805
## 123  0.58236519
## 124 -2.75406818
## 125  0.18703400
## 126  1.26739882
## 127  0.27142497
## 128 -2.74013261
## 129  1.36456091
## 130 -0.55820367
## 131 -2.73443140
## 132 -0.55820367
## 133  1.36456091
## 134 -0.55820367
## 135 -4.41443147
## 136  1.11350315
## 137  0.27142497
## 138  1.82207354
## 139 -3.78266427
## 140  1.73646579
## 141 -1.98034152
## 142  0.31381047
## 143 -1.63533466
## 145  0.31942671
## 201  1.17208118
## 202  0.60548770
## 203 -1.63533466
## 204  0.82996151
## 205  0.18703400
## 206 -0.55820367
## 207  2.24127645
## 208  0.03645967
## 209  0.20918236
## 210 -1.56581860
## 211  0.31381047
## 213  1.26739882
## 601 -1.83949657
## 602  0.60548770
## 603 -1.41743848
## 604  0.27142497
## 605 -0.76011560
## 606 -5.48931240
## 607  0.60503139
## 608  0.39872805
## 609  0.60548770
## 610  0.39872805
## 611  1.11350315
## 612 -1.75735366
## 613  0.20918236
## 614  0.19105528
## 615  0.39872805
## 616  0.60548770
## 617  0.60548770
## 618  1.36456091
## 619 -0.28405722
## 620  0.27142497
## 621  0.03645967
## 622 -0.32860690
## 624 -0.50252279
## 625  0.39872805
##
## with conditional variances for "ID"
```

```
fixed.effects(h.GLMM1)
```

```
##           (Intercept)             baseline1         AGEGROUP25-34 
##            0.19521497           -2.77610394            2.25650715 
##           AGEGROUP35+ treatmentIntervention                 month 
##            1.98228759            3.41324756            0.03718025
```

$$g(E(Y_{ij}|b_i)) = X_{ij}\beta + Z_{ij}b_i + \epsilon_{ij}$$

Interpretation:

- Only the time variable *month* can be interpreted in this experiement design.
- *month*: For patients within the same age group and same treatment group and same baseline rating, the log odds ratio of good health rating for one month increase is 0.03718.
- bashealthPoor, agegroup25-34, agegroup35+ cannot be interpreted, because for the same subject, these characteristics are not interchangeable, and there is no point in comparing it to different levels.

The difference is that the GEE model focuses on modelling Population-average while the GLMM model focuses on the specific individual and some variables cannot be changed for an individual.