Homework 8

Xinyi Lin

4/21/2019

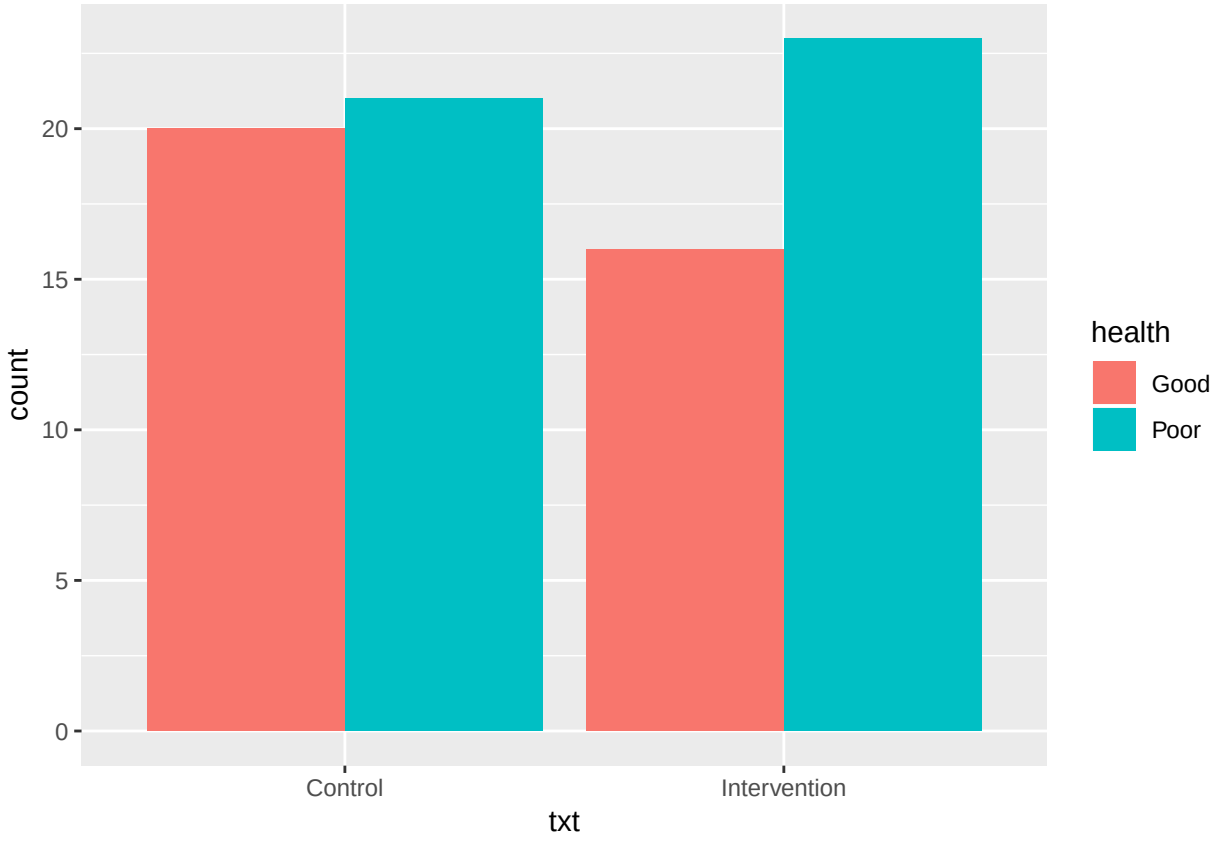
```
library(tidyverse)
library(ggplot2)
library(gee)
library(lme4)
```

```
library(readxl)
health <- read_excel("HEALTH.xlsx") %>%
  janitor::clean_names() %>%
  mutate(health = as.factor(health))
```

Question 1

Draw a plot about numbers of 'Good' health group and 'Poor' health group in control group and intervention group when the time equals to 1.

```
health %>%
  filter(time == 1) %>%
  ggplot(aes(x=txt, fill=health)) +
  geom_bar(position="dodge") +
  xlab("txt") + labs(fill="health")
```



According to the plot, ratios of people whose self-rating levels are good or poor separetly in control group are slightly different from those in intervention group.

Fit a GLM model to find out whether group assignment is a significant variable when predicting self-rated level.

```
health_random = health %>%
  filter(time == 1)
glm1 = glm(health~txt, family = binomial, health_random)
summary(glm1)
```

```
##
## Call:
## glm(formula = health ~ txt, family = binomial, data = health_random)
##
## 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
```

According to the model, we can find that `txt` variable is not significant so there is no significant relationship between randomized group assignment and participants health self-rating.

Question 2

Fit the GEE model with unstructure correlation structure.

```
subset <- subset(health, time > 1)
# make month 0 as another covariate: baseline
base = subset(health, time == 1) %>%
  select(id, health)
colnames(base) = c("id", "baseline")
new_subset = merge(subset,base,by="id",all=F)
new_subset$nhealth <- as.numeric(new_subset$health == "Good") # 1=Good, 0=Poor
new_subset$time = plyr::mapvalues(new_subset$time, from = c(2,3,4), to = c(3,6,12))

#head(new_subset)

gee.fit = gee(nhealth~txt+time+agegroup+baseline,
                data = new_subset, family = "binomial",
                id = id ,corstr = "unstructured",
                scale.fix = TRUE, scale.value = 1)
```

```
##     (Intercept) txtIntervention            time   agegroup25-34
```

```
##      0.18528086      1.99669985      0.02536275      1.19749448
##     agegroup35+    baselinePoor
##      1.39742621     -1.71063852
```

```
summary(gee.fit)
```

```
##
##  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 = nhealth ~ txt + time + agegroup + baseline, id = id,
##     data = new_subset, 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.   Robust z
## (Intercept)      0.12457924 0.47137316  0.2642901  0.51374172  0.2424939
## txtIntervention  2.10225898 0.48779381  4.3097286  0.53777951  3.9091467
## time             0.03243343 0.03665686  0.8847848  0.04755408  0.6820326
## agegroup25-34    1.35250468 0.48130172  2.8100973  0.50420159  2.6824681
## agegroup35+      1.42052166 0.79781620  1.7805124  0.78372968  1.8125148
## baselinePoor    -1.81418056 0.48958528 -3.7055456  0.50961334 -3.5599158
##
## 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
```

Interpretation:

Intercept: The log odd of self-rating of health status as good is 0.12 on average for sub-population in control, age 15-24 group at the time of randomization with Good baseline.

β_{txt} : The log odd ratio of self-rating of health status as good between intervention group vs control group is 2.10 on average for sub-population with same age group, baseline in same time.

β_{time} : The log odd ratio of self-rating of health status as good is 0.03 on average with one unit changes in time for sub-population with same age group, baseline and treatment group.

$\beta_{age25-34}$: The log odd ratio of self-rating of health status as good between 25-34 age group vs 15-24 age group is 1.35 on average for sub-population with same treatment group, baseline in same time.

β_{age35+} : The log odd ratio of self-rating of health status as good between 35+ age group vs 15-24 age group

is 1.42 on average for sub-population with same treatment group, baseline in same time.

β_{base} : The log odd ratio of self-rating of health status as good between poor baseline group vs good baseline group is -1.81 on average for sub-population with same treatment group and age group in same time.

Question 3

Fit GLMM model. The model is as following:

$$logit(E(Y_{ij}|b_i)) = (b_i + \beta_1) + X_{ij}^T\beta$$

```
glmm.fit <- glmer(nhealth ~ baseline + txt + time + agegroup + (1 | id),
                  family = 'binomial', data = new_subset)
summary(glmm.fit)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: nhealth ~ baseline + txt + time + agegroup + (1 | id)
##    Data: new_subset
##
##      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
## baselinePoor    -2.77610    0.98381  -2.822  0.00478 **
## txtIntervention  3.41325    1.07267   3.182  0.00146 **
## time             0.03718    0.06933   0.536  0.59176
## agegroup25-34    2.25651    1.00877   2.237  0.02529 *
## agegroup35+      1.98229    1.38118   1.435  0.15123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) bslnPr txtInt time   a25-34
## baselinePor -0.374
## txtIntrvntn -0.256 -0.449
## time        -0.472 -0.016  0.047
## agegrp25-34 -0.319 -0.379  0.395  0.007
## agegroup35+ -0.195 -0.274  0.206 -0.007  0.390
```

Interpretation:

β_0 : The log odd of self-rating of health status as good is 0.20 on average across all subjects with baseline equals to good, in control group and 15-24 age group at the time of randomization.

β_{base} : Cannot interpret.

β_{txt} : Cannot interpret.

β_{time} : The log odd ration of self-rating of health status as good is 0.04 with one unit changes in time for the same subject.

$\beta_{age25-34}$: Cannot interpret.

β_{age35+} : Cannot interpret.

Difference:

For the GEE model, it considers the situation of sub-population, so its interpretation is based on sub-population and all coefficients can be interpreted. However, for the GLMM mdoel, it considers specific individual and some of its coefficients.