P8131-HW8-yz4184

Yunlin Zhou

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

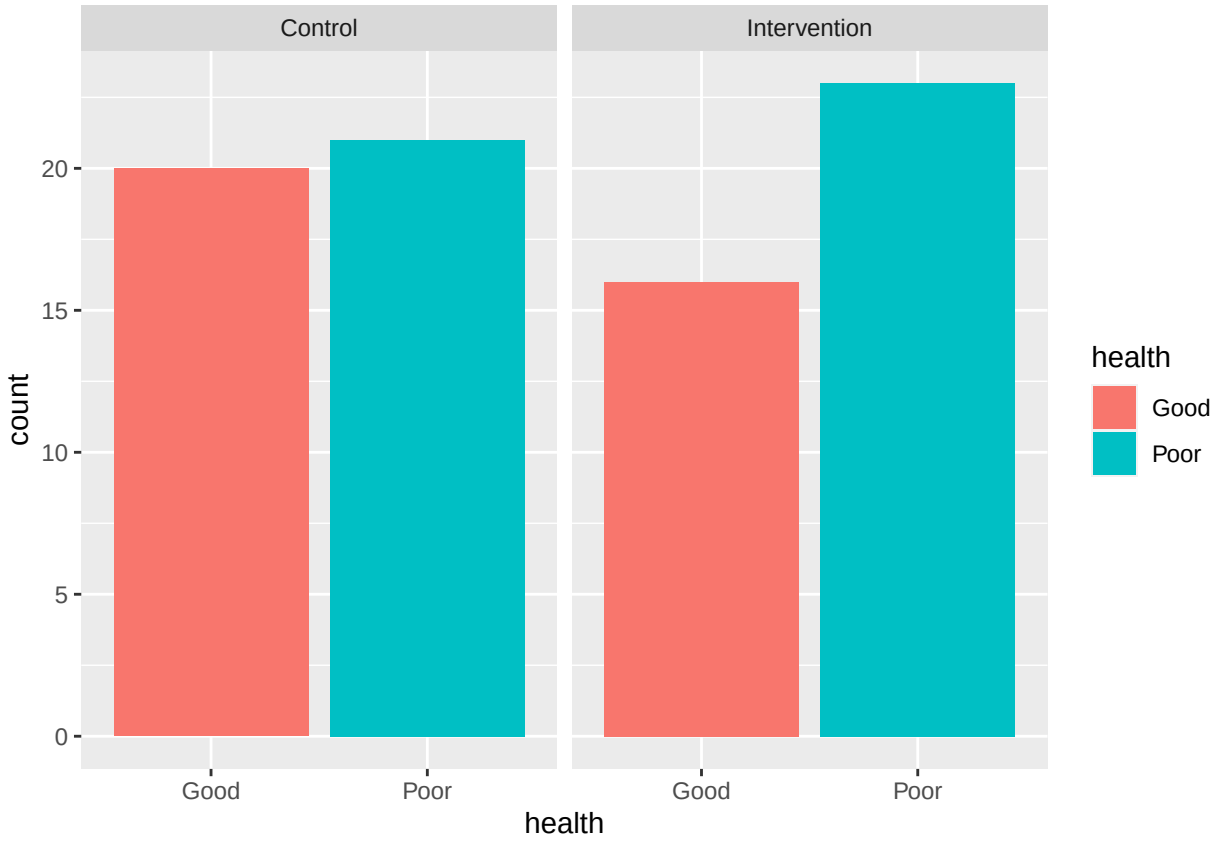
```
# Import and clean the data
health_df = readxl::read_xlsx("./HW8-HEALTH.xlsx")%>%
  janitor::clean_names()%>%
  drop_na()%>%
  mutate(txt = as.factor(txt),
         health = as.factor(health),
         agegroup = as.factor(agegroup),
         time = as.integer(time))
```

Part A

First, we plot the count of participants health self-rating counts in different treatment groups.

```
health_random = health_df %>%
  filter(time == 1)

ggplot(data = health_random, aes(x= health, fill = health)) +
  geom_bar() +
  facet_grid(cols = vars(txt))
```



At the time of randomization, the control group doesn't show significant difference in health self-rating. However, the intervention group shows more poor self-ratings than good self-ratings.

Then, we would like to use logistic model to test the relationship between randomized group assignment and participants health self-rating.

```
glm = glm(health ~ txt,
          family = binomial(link = "logit"),
          data = health_random)

summary(glm)
```

```
##
## Call:
## glm(formula = health ~ txt, family = binomial(link = "logit"),
##     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
```

The odds ratio of having good health is expected to be 1.3690476 for intervention group compared to control group.

The p value for the coefficient is 0.486 which is greater than 0.05, which indicates strong evidence for the null hypothesis that there is no correlation between treatment group assignment and health status at the time of randomization.

Part b

```
health_follow = health_df %>%
  mutate(health = as.numeric(health == "Good"))%>%
  filter(!id %in% names(which(table(health_df$id) == 1)))

resp <- subset(health_follow, time != 1)

resp$baseline <- rep(subset(health_follow, time == 1)$health, as.numeric(table(resp$id)))

resp$nhealth <- as.numeric(resp$health == "Good")

gee = gee(health ~ baseline + txt + time + agegroup,
          family = "binomial",
          corstr = "unstructured",
          id = id,
          scale.fix = FALSE,
          data = resp)
```

```
##     (Intercept)        baseline txtIntervention            time   agegroup25-34
##      -1.7414839       1.7112931       1.9977806       0.1321222       1.1958638
##     agegroup35+
##       1.3954271
```

```
summary(gee)
```

```
##
##  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 ~ baseline + txt + time + agegroup, id = id,
##     data = resp, family = "binomial", corstr = "unstructured",
##     scale.fix = FALSE)
##
## Summary of Residuals:
##         Min          1Q      Median          3Q         Max
## -0.98120150 -0.18801168  0.09128879  0.17516123  0.83424138
##
##
## Coefficients:
##                   Estimate Naive S.E.    Naive z Robust S.E.  Robust z
## (Intercept)     -1.9220068  0.7873221 -2.4411949   0.7369212 -2.608158
## baseline         1.8144864  0.6033350  3.0074276   0.5104410  3.554743
## txtIntervention  2.0995031  0.6008738  3.4940832   0.5379270  3.902951
## time             0.1530083  0.2017530  0.7583941   0.2107268  0.726098
## agegroup25-34    1.3509848  0.5930043  2.2782040   0.5038608  2.681266
## agegroup35+      1.4116600  0.9825238  1.4367693   0.7864438  1.794992
##
## Estimated Scale Parameter:  1.516997
## Number of Iterations:  5
##
## Working Correlation
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.1743007 0.5809889
## [2,] 0.1743007 1.0000000 0.2049833
## [3,] 0.5809889 0.2049833 1.0000000
```

The odds ratio of participants self-rating "good" compared to "poor" is 6.137923, at at the time of randomization, if take average among all measurements and all subjects within the same subgroup.

The odds ratio of participants self-rating "good" compared to "poor" is 8.1621131, compare "intervention" to "control" treatment group, if take average among all measurements and all subjects within the same subgroup.

The odds ratio of participants self-rating "good" compared to "poor" is 1.1653347, for the follow-up time after the time of randomization., if take average among all measurements and all subjects within the same subgroup.

The odds ratio of participants self-rating "good" compared to "poor" is 3.8612263, compare 25-34 age group to 15-24 age group, if take average among all measurements and all subjects within the same subgroup.

The odds ratio of participants self-rating "good" compared to "poor" is 3.8612263, compare 35+ age group to 15-24 age group, if take average among all measurements and all subjects within the same subgroup.

Part c

```
glmm = glmer(health ~ baseline + txt + time + agegroup + (1 | id),
             family = "binomial",
             data = resp)

summary(glmm)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: health ~ baseline + txt + time + agegroup + (1 | id)
##    Data: resp
##
##      AIC      BIC   logLik deviance df.resid
##    184.8    207.9    -85.4    170.8      192
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.5391 -0.2367  0.1427  0.2909  1.8719
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  id     (Intercept) 5.765    2.401
## Number of obs: 199, groups:  id, 78
##
## Fixed effects:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.9240     1.3015  -2.247  0.02467 *
## baseline          2.7813     0.9874   2.817  0.00485 **
## txtIntervention   3.4231     1.0780   3.176  0.00150 **
## time              0.2021     0.3090   0.654  0.51298
## agegroup25-34     2.2587     1.0128   2.230  0.02573 *
## agegroup35+       1.9803     1.3853   1.430  0.15286
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) baseln txtInt time   a25-34
## baseline    -0.526
## txtIntrvntn -0.542  0.450
## time        -0.680  0.034  0.068
## agegrp25-34 -0.514  0.380  0.396  0.022
## agegroup35+ -0.340  0.275  0.206 -0.002  0.390
```

The odds ratio of participants self-rating "good" compared to "poor" is 16.1399893, at at the time of randomization, if take average among all measurements and all subjects within the same subgroup.

The odds ratio of participants self-rating "good" compared to "poor" is 30.6643272, compare "intervention" to "control" treatment group, if take average among all measurements and all subjects within the same subgroup.

The odds ratio of participants self-rating "good" compared to "poor" is 1.2239704, for the follow-up time after the time of randomization., if take average among all measurements and all subjects within the same subgroup.

The odds ratio of participants self-rating "good" compared to "poor" is 9.5706392, compare 25-34 age group to 15-24 age group, if take average among all measurements and all subjects within the same subgroup.

The odds ratio of participants self-rating "good" compared to "poor" is 7.2449161, compare 35+ age group to 15-24 age group, if take average among all measurements and all subjects within the same subgroup.