HW8

Zirui Zhang

2023-04-19

```
library(tidyverse)
library(dplyr)
library(readxl)
library(sqldf)
library(gee)
library(lme4)
library(nlme)
```

```
# data import
health = read_xlsx("./HW8-HEALTH.xlsx") %>%
  janitor::clean_names() %>%
  mutate(health = as.numeric(health == "Good"),
         time = ifelse(time==2, 3,
                ifelse(time==3, 6,
                ifelse(time==4, 12, 1))))
```

Question (a)

```
group = health %>%
  mutate(txt = as.numeric(txt == "Intervention")) %>%
  filter(time == 1)

lg = glm(txt ~ health, data = group, family = binomial(link = 'logit'))
summary(lg)
```

```
##
## Call:
## glm(formula = txt ~ health, family = binomial(link = "logit"),
##     data = group)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.216  -1.216  -1.084   1.139   1.274
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.09097    0.30182   0.301    0.763
## health      -0.31412    0.45122  -0.696    0.486
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 110.85  on 79  degrees of freedom
## Residual deviance: 110.37  on 78  degrees of freedom
## AIC: 114.37
##
## Number of Fisher Scoring iterations: 3
```

$$log\frac{\pi}{1-\pi} = 0.091 - 0.314 * health$$

When a patient's health changes from "poor" to "good", the odds of he/she assigned to the intervention group decrease by 0.2696. However, as the p-value is greater than 0.05, the relationship is not significant here. Being in poor or good health at baseline wouldn't affect the group of which the patient would be assigned into.

Question (b)

```
# data manipulation
base = health %>%
  filter(time == 1)%>%
  select(id, health) %>%
  rename(baseline = health)
# make time 1 as variable baseline
gee_df = sqldf("
      SELECT *
      FROM health
      LEFT JOIN base
      USING(id)
      ") %>%
  filter(time != 1)
# fit gee model
gee = gee(health ~ as.factor(baseline) + txt + as.factor(time) + agegroup, data=gee_df, family="binomial
```

```
##         (Intercept) as.factor(baseline)1      txtIntervention
##          -1.5199450            1.7192117            2.0042708
##    as.factor(time)6    as.factor(time)12        agegroup25-34
##           0.2575654            0.2366989            1.1968673
##         agegroup35+
##           1.3958656
```

```
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 ~ as.factor(baseline) + txt + as.factor(time) +
##     agegroup, id = id, data = gee_df, family = "binomial", corstr = "unstructured",
##     scale.fix = FALSE)
##
## Summary of Residuals:
##         Min          1Q      Median          3Q         Max
## -0.97980130 -0.20060701  0.09442344  0.18344971  0.83995062
##
##
## Coefficients:
##                        Estimate Naive S.E.    Naive z Robust S.E.  Robust z
## (Intercept)          -1.6578607  0.6014505 -2.7564377   0.4533989 -3.656517
## as.factor(baseline)1  1.8164161  0.5978966  3.0380103   0.5113296  3.552339
## txtIntervention       2.1022271  0.5954429  3.5305269   0.5362768  3.920041
## as.factor(time)6      0.2753559  0.4747047  0.5800572   0.3368572  0.817426
## as.factor(time)12     0.2863563  0.4083916  0.7011809   0.4161352  0.688133
## agegroup25-34         1.3345925  0.5860828  2.2771400   0.5043829  2.645991
## agegroup35+           1.4112905  0.9740226  1.4489299   0.7855584  1.796544
##
## Estimated Scale Parameter:  1.486693
## Number of Iterations:  4
##
## Working Correlation
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.1794182 0.5602284
## [2,] 0.1794182 1.0000000 0.2104116
## [3,] 0.5602284 0.2104116 1.0000000
```

```
exp(gee$coefficients)
```

```
##          (Intercept) as.factor(baseline)1      txtIntervention
##            0.1905462            6.1497789            8.1843772
##     as.factor(time)6    as.factor(time)12        agegroup25-34
##            1.3169993            1.3315669            3.7984479
##          agegroup35+
##            4.1012448
```

Interpretation:

- Baseline: On average within a population, holding other variables fixed, the odds ratio of reporting good health would be 6.15 between subpopulations reporting good health compared to poor health at baseline;
- Treatment group: On average within a population, holding other variables fixed, the odds ratio of reporting good health would be 8.18 between subpopulations in intervention group compared to control group;
- 6 month: On average within a population, holding other variables fixed, the odds ratio of reporting good health would be 1.32 between reports of subpopulations at 6 months compared to at 3 months;
- 12 month: On average within a population, holding other variables fixed, the odds ratio of reporting good health would be 1.33 between reports of subpopulations at 12 months compared to at 3 months;

- Age group 25-34: On average within a population, holding other variables fixed, the odds ratio of reporting good health would be 3.80 between subpopulations of age 25-34 compared to those of age 15-24;
- Age group 35+: On average within a population, holding other variables fixed, the odds ratio of reporting good health would be 3.80 between subpopulations of age 35+ compared to those of age 15-24;

Question (c)

```
# fit glmm model
glmm = glmer(health ~ time + txt + agegroup + (1|id), family = "binomial", data = health)
summary(glmm)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: health ~ time + txt + agegroup + (1 | id)
##    Data: health
##
##      AIC      BIC   logLik deviance df.resid
##    328.6    350.3   -158.3    316.6      273
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.6168 -0.5338  0.3268  0.5211  2.0710
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  id     (Intercept) 2.386    1.545
## Number of obs: 279, groups:  id, 80
##
## Fixed effects:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.92699    0.46058  -2.013   0.0441 *
## time             0.14397    0.04613   3.121   0.0018 **
## txtIntervention  1.21265    0.48585   2.496   0.0126 *
## agegroup25-34    0.76899    0.50895   1.511   0.1308
## agegroup35+      0.34509    0.83832   0.412   0.6806
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) time   txtInt a25-34
## time        -0.436
## txtIntrvntn -0.510  0.064
## agegrp25-34 -0.507  0.004 -0.027
## agegroup35+ -0.317 -0.017  0.018  0.289
```

```
exp(fixed.effects(glmm))
```

```
##     (Intercept)            time txtIntervention   agegroup25-34     agegroup35+
##       0.3957412       1.1548545       3.3623948       2.1575760       1.4121109
```

- Time: On average for an individual, holding other variables, when time increases by 1 month, we would expect to see a 15.5% increase in the odds of reporting good health;
- Treatment group: On average for an individual, holding other variables fixed, when treatment group changes from control to intervention, we would expect to see a 236.2% increase in the odds of reporting good health;
- Age group 25-34: On average for an individual, holding other variables fixed, when age group changes from 15-24 to 25-34, we would expect to see a 115.8% increase in the odds of reporting good health;
- Age group 35+: On average for an individual, holding other variables fixed, when age group changes from 25-34 to 35+, we would expect to see a 41.2% increase in the odds of reporting good health;
- Random effects:
$$b_i \sim N(0, \sigma^2)$$
where
$$\sigma = 1.545$$

Difference between interpretations:

The interpretation of GEE models focus on within population level change while the interpretation of GLMM models focus on individual level change.