BM2_HW8

Yixiao Sun

2024-04-11

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(summarytools)
```

```
##
## Attaching package: 'summarytools'
##
## The following object is masked from 'package:tibble':
##
##     view
```

```
library(ggplot2)
library(bayesQR)
library(readxl)
library(gee)
health_data <- read_excel("~/Desktop/P8130_Biostatistical Method/BM2_HW8/HW8-HEALTH.xlsx")
View(health_data)
dim(health_data)
```

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

```
head(health_data)
```

```
## # 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
```

```
health_data$HEALTH2<-as.numeric(health_data$HEALTH == "Good")
health_data$AGEGROUP<-as.factor(health_data$AGEGROUP)
health_data2 <- subset(health_data,health_data$TIME == "1")
```

a)

```
logit.fit <- glm(formula = HEALTH2 ~ TXT,family = binomial,data = health_data2)
summary(logit.fit)
```

```
##
## Call:
## glm(formula = HEALTH2 ~ TXT, family = binomial, data = health_data2)
##
## 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
```

Based on the results of the logistic model, we conclude that the TXT are not significant for having a p_value of 0.486, which is bigger than alpha 0.05, so we do not reject the null hypothesis. Therefore, there arent significant relationship between the randomization and how the patients self identify their health status.

b)

```
health_data3 <- health_data %>%
    group_by(ID) %>%
    mutate(baseline = HEALTH[TIME == "1"],
           TIME = case_match(TIME,
                             2 ~ 3,
                             3 ~ 6,
                             4 ~ 12)) %>% ungroup(ID) %>% subset(TIME > "1")
health_data3$nstat <- as.numeric(health_data3$HEALTH == "Good")
gee_model <- gee(formula = nstat ~ baseline + TXT + TIME + AGEGROUP,
                 data = health_data3,
```

```
                    id = ID,
                    family = binomial,
                    corstr = "unstructured",scale.fix = T,scale.value = 1)
```

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

## running glm to get initial regression estimate

##     (Intercept)    baselinePoor TXTIntervention            TIME   AGEGROUP25-34
##      0.18528086     -1.71063852      1.99669985      0.02536275      1.19749448
##     AGEGROUP35+
##      1.39742621
```

```
summary(gee_model)
```

```
##
##  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 = nstat ~ baseline + TXT + TIME + AGEGROUP, id = ID,
##     data = health_data3, family = binomial, corstr = "unstructured",
##     scale.fix = T, 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
## baselinePoor    -1.81418056 0.48958528 -3.7055456  0.50961334 -3.5599158
## 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
##
## 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
```

Without randomization, based on the model, the odds ratio of self rating as good and the baseline as good compared with poor baseline is estimated as 1.814. And the odds ratio of having self rating as good for those in the treatment group compared with those in the control group is estimaated as 2.10225. The odds ratio of having self rating for every unit increase in after randomization month is 0.03243. And the odds ratio of having self rating as good for those who are in the age group of 25-34 compared with those in the age group of 15-24 is estimated as 1.3525. Finally, the odds ratio of having self rating as good for those who are in the age group of 35 and older compared with those in the age group of 15-24 is estimated as 1.42052.

c)

```
library(lme4)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
library(Matrix)
library(nlme)
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:lme4':
##
##     lmList
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
GLMEM <- glmer(nstat ~ baseline + TXT + TIME + AGEGROUP + (1|ID),
               data = health_data3,
               family = binomial)
summary(GLMEM)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: nstat ~ baseline + TXT + TIME + AGEGROUP + (1 | ID)
##    Data: health_data3
##
##      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    1.07268   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    1.38119   1.435  0.15123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) bslnPr TXTInt TIME   AGEGROUP2
## baselinePor -0.374
## TXTIntrvntn -0.256 -0.449
## TIME        -0.472 -0.016  0.047
## AGEGROUP25- -0.319 -0.379  0.395  0.007
## AGEGROUP35+ -0.195 -0.274  0.206 -0.007  0.390
```

The odds ratio of having self rating for every unit increase in after randomization month is 0.03718 for the same subject. The main difference for the two models we made in (B) and (C) is that for this linear mixed model mainly focus on the individual level, but the GEE model made in question 2 are mainly focus on the population mean.