Homework 8

Yuki Joyama

2024-04-07

```
# data prep
df = read_excel("HW8-HEALTH.xlsx") |>
  janitor::clean_names() |>
  rename(trt = txt)

head(df)
```

```
## # A tibble: 6 x 5
##      id  time trt          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
```

(a)

```
# limit data to time == 1
df_base = df |>
  filter(time == 1)

head(df_base)
```

```
## # A tibble: 6 x 5
##      id  time trt          health agegroup
##   <dbl> <dbl> <chr>        <chr>  <chr>
## 1   101     1 Intervention Good   15-24
## 2   102     1 Control      Poor   15-24
## 3   103     1 Control      Good   25-34
## 4   104     1 Intervention Good   25-34
## 5   105     1 Intervention Poor   15-24
## 6   106     1 Control      Poor   25-34
```

```
# summarize randomized group assignment and health self_rating at the time of randomization
theme_gtsummary_journal(journal = "nejm")
```

```r
df_base |>
  select(trt, health) |>
  tbl_summary(
    by = trt,
    statistic = list(
      all_categorical() ~ "{n} ({p}%)"
    ),
    digits = all_continuous() ~ 1,
    label = list(
      health ~ "Health"
    )
  ) |>
  # modify_caption("Table 1: Baseline Characteristics") |>
  as_flex_table()
```

Characteristic	**Control**, N = 41[1]	**Intervention**, N = 39[1]
Health		
Good	20 (49%)	16 (41%)
Poor	21 (51%)	23 (59%)

[1] n (%)

We can see that the Intervention group exhibits a slightly higher count of individuals who reported "poor" in their self-ratings compared to the Control group.

```r
# chi-square test
contingency_table <- table(df_base$trt, df_base$health)

chisq.test(contingency_table)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  contingency_table
## X-squared = 0.22287, df = 1, p-value = 0.6369
```

The Pearson's chi-squared test (p-value = 0.6369 > 0.05) indicates that there is no statistically significant difference in health self-rating between two treatment groups at the time of randomization.

(b)

To analyze the participants' self-rated level of health across all study follow-up visits (except for the time of randomization), I will fit a GEE model with health self-rating at the baseline, treatment group, month post randomization, and age group as predictors.
The correlation structure is set as `unstructured` in this model.

```r
# data prep; make time 1 as another covariate
df_fu = df |>
  group_by(id) |>
  mutate(baseline = health[time == 1]) |> # record the baseline health ratings
  mutate(nstat = ifelse(health == "Poor", 0, 1)) |> # poor: 0, good: 1
  filter(time != 1) |>
  ungroup() |>
  arrange(id) # arrange the data based on the group variable

head(df_fu)
```

```
## # A tibble: 6 x 7
##      id  time trt          health agegroup baseline nstat
##   <dbl> <dbl> <chr>        <chr>  <chr>    <chr>    <dbl>
## 1   101     2 Intervention Good   15-24    Good         1
## 2   101     3 Intervention Good   15-24    Good         1
## 3   101     4 Intervention Good   15-24    Good         1
## 4   102     2 Control      Poor   15-24    Poor         0
## 5   102     3 Control      Poor   15-24    Poor         0
## 6   102     4 Control      Poor   15-24    Poor         0
```

```r
# fit logistic GEE with unstructured correlation structure
fit.gee = gee(
  nstat ~ baseline + trt + as.factor(time) + agegroup,
  data = df_fu, family = "binomial", id = id,
  corstr = "unstructured", scale.fix = FALSE
)
```

```
##      (Intercept)     baselinePoor  trtIntervention as.factor(time)3
##        0.1992666       -1.7192117        2.0042708        0.2575654
## as.factor(time)4    agegroup25-34      agegroup35+
##        0.2366989        1.1968673        1.3958656
```

```r
summary(fit.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 = nstat ~ baseline + trt + as.factor(time) + agegroup,
##     id = id, data = df_fu, family = "binomial", corstr = "unstructured",
##     scale.fix = FALSE)
##
## Summary of Residuals:
##         Min          1Q      Median          3Q         Max
## -0.97980441 -0.20059815  0.09443036  0.18342395  0.83995790
```

```
##
##
## Coefficients:
##                    Estimate Naive S.E.    Naive z Robust S.E.   Robust z
## (Intercept)       0.1585228  0.5584402  0.2838671   0.4525895  0.3502574
## baselinePoor     -1.8164377  0.5979339 -3.0378569   0.5113338 -3.5523518
## trtIntervention   2.1021788  0.5954745  3.5302585   0.5362723  3.9199838
## as.factor(time)3  0.2753548  0.4747121  0.5800458   0.3368632  0.8174082
## as.factor(time)4  0.2863899  0.4082589  0.7014909   0.4161677  0.6881598
## agegroup25-34     1.3347971  0.5861276  2.2773149   0.5043892  2.6463636
## agegroup35+       1.4111223  0.9740514  1.4487144   0.7856000  1.7962350
##
## Estimated Scale Parameter:  1.486721
## Number of Iterations:  5
##
## Working Correlation
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.1794168 0.5605611
## [2,] 0.1794168 1.0000000 0.2105039
## [3,] 0.5605611 0.2105039 1.0000000
```

```
# OR
exp(fit.gee$coef)
```

```
##      (Intercept)     baselinePoor  trtIntervention as.factor(time)3
##         1.171779         0.162604         8.183982         1.316998
## as.factor(time)4    agegroup25-34      agegroup35+
##         1.331612         3.799225         4.100555
```

In our analysis, participants with a `Poor` health self-rating at baseline were found to have 0.16 times the odds of reporting a `Good` self-rating compared to those with a `Good` health self-rating at baseline, holding other variables constant. Additionally, all else being equal, the intervention group exhibited 8.18 times the odds of reporting a `Good` self-rating. Moreover, individuals in the age group 25-34 had 3.8 times the odds of reporting a `Good` self-rating compared to those in the age group 15-24.
We should note that the covariates `time = 3`, `time = 4`, and `agegroup = 35+` do not have statistically significant influence on the response variable given the robust z values ($z < 1.96$).

(c)

Here I will fit a generalized linear mixed effects model with subject-specific random intercepts.

```
# fit glmm with subject-specific random intercepts
fit.glmm = glmer(
  nstat ~ baseline + trt + as.factor(time) + agegroup + (1 | id),
  data = df_fu, family = "binomial", nAGQ = 0
)

summary(fit.glmm)
```

```
## Generalized linear mixed model fit by maximum likelihood (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 0) [glmerMod]
```

```
##  Family: binomial  ( logit )
## Formula: nstat ~ baseline + trt + as.factor(time) + agegroup + (1 | id)
##    Data: df_fu
##
##      AIC      BIC   logLik deviance df.resid
##    189.9    216.2    -86.9    173.9      191
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.2579 -0.3220  0.2358  0.3537  1.7935
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  id     (Intercept) 3.522    1.877
## Number of obs: 199, groups:  id, 78
##
## Fixed effects:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         0.1262     0.6576   0.192 0.847812
## baselinePoor       -1.8678     0.6515  -2.867 0.004147 **
## trtIntervention     2.2710     0.6470   3.510 0.000448 ***
## as.factor(time)3    0.3557     0.4939   0.720 0.471458
## as.factor(time)4    0.2863     0.5472   0.523 0.600869
## agegroup25-34       1.4720     0.6753   2.180 0.029279 *
## agegroup35+         1.2985     1.0629   1.222 0.221867
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) bslnPr trtInt as.()3 as.()4 a25-34
## baselinePor -0.462
## trtIntrvntn -0.373 -0.125
## as.fctr(t)3 -0.317 -0.030  0.038
## as.fctr(t)4 -0.281  0.023 -0.002  0.416
## agegrp25-34 -0.438 -0.067  0.051 -0.018 -0.047
## agegroup35+ -0.223 -0.148  0.034 -0.016 -0.028  0.298
```

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

```
##      (Intercept)     baselinePoor  trtIntervention as.factor(time)3
##        1.1345099        0.1544695        9.6887266        1.4271410
## as.factor(time)4    agegroup25-34      agegroup35+
##        1.3314268        4.3577500        3.6636243
```

Similar to the GEE model, the covariates `time = 3`, `time = 4`, and `agegroup = 35+` do not have statistically significant influence on the response variable in this model.
The fixed effects tells us that participants with a `Poor` health self-rating at baseline were found to have 0.15 times the odds of reporting a `Good` self-rating compared to those with a `Good` health self-rating at baseline, holding other variables and random effects constant. All else being equal, the intervention group exhibited 9.69 times the odds of reporting a `Good` self-rating than the control group. The result also indicates that the individuals in the age group 25-34 had 4.36 times the odds of reporting a `Good` self-rating compared to those in the age group 15-24, with other variables and random effects unchanged.

The subject-specific random intercepts are as follows:

```r
random.effects(fit.glmm)
```

```
## $id
##     (Intercept)
## 101  0.46910205
## 102 -0.88027986
## 103  0.75821530
## 104  0.15249226
## 105 -0.03247029
## 106  1.70902511
## 107  1.24488498
## 109  0.15249226
## 110  1.48177656
## 111  0.15249226
## 112 -1.87893092
## 113 -0.68204013
## 114  0.60196325
## 116  0.66668026
## 117 -1.28801476
## 118 -0.73576015
## 119  0.47080535
## 120 -0.76527235
## 121 -0.68204013
## 122  0.60196325
## 123  0.66737503
## 124 -2.12802630
## 125  0.31593641
## 126  0.57325132
## 127  0.46910205
## 128 -1.79817957
## 129  1.26281205
## 130 -0.68204013
## 131 -1.59282890
## 132 -0.68204013
## 133  1.26281205
## 134 -0.68204013
## 135 -3.09953737
## 136  0.91737499
## 137  0.46910205
## 138  1.45066394
## 139 -2.06672034
## 140  1.48177656
## 141 -1.64392713
## 142  0.47080535
## 143 -1.35382068
## 145  0.16866274
## 201  1.04804842
## 202  0.75821530
## 203 -1.35382068
## 204  0.76038553
## 205  0.31593641
## 206 -0.68204013
## 207  1.81204898
```

```
## 208  0.15249226
## 209  0.36013961
## 210 -0.76527235
## 211  0.47080535
## 213  0.57325132
## 601 -0.91816895
## 602  0.75821530
## 603 -1.02148717
## 604  0.46910205
## 605 -0.88027986
## 606 -3.42670881
## 607  0.53212877
## 608  0.60196325
## 609  0.75821530
## 610  0.60196325
## 611  0.91737499
## 612 -1.19539381
## 613  0.36013961
## 614  0.25106958
## 615  0.60196325
## 616  0.75821530
## 617  0.75821530
## 618  1.26281205
## 619 -0.32987773
## 620  0.46910205
## 621  0.15249226
## 622 -0.37770711
## 624 -0.45310038
## 625  0.60196325
##
## with conditional variances for "id"
```

The primary distinction between the GEE model and the GLMM lies in their focus: the GEE model provides insights into the **population-averaged** odds ratio/log odds of the response variable, while the GLMM with subject-specific random intercepts allows for the examination of individual variations. In the GLMM, participants possess both random intercepts, which vary among individuals, and fixed intercepts, which remain consistent across the population. Thus, we can infer the **subject-specific** odds ratio/log odds using the GLMM.