Homework 8

Yihan Feng

The data (HEALTH.xlsx) are from a randomized, controlled trial among women of childbearing age to evaluate the effects of an educational intervention. One response variable of interest is the participants self-rating of health status as either good or poor. The researchers would like to assess the effect of the intervention on self-rated health across the follow-up period, as well as whether these effects are influenced by the mothers age. There are n = 80 women enrolled in this trial. These data were measured at 4 points in time: randomization, 3 months, 6 months, and 12 months post-randomization.

```
health.df = read_excel("./HW8-HEALTH.xlsx") %>%
  janitor::clean_names() %>%
  rename(trt = txt) %>%
  mutate(trt = as.factor(trt),
         health = as.numeric(health == "Good"),
         time = as.integer(time))
health.df1 = health.df %>%
  filter(!id %in% names(which(table(health.df$id) == 1))) # remove the participants with randomization
```

(a) Evaluate the bivariate, cross-sectional relationship between randomized group assignment and participants health self-rating at the time of randomization. Interpret and discuss these findings.

```
health.df.a = health.df %>%
  filter(time == 1)
glm = glm(health ~ trt,
                  data = health.df.a,
                  family = binomial(link = "logit"))
summary(glm)
```

```
##
## Call:
## glm(formula = health ~ trt, family = binomial(link = "logit"),
##     data = health.df.a)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.157  -1.157  -1.028   1.198   1.335
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.04879    0.31244  -0.156    0.876
## trtIntervention -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 self-reporting “good” health status at randomization (baseline) is 0.73, for intervention group vs. control group. However, the p value for the coefficient is $0.486 > 0.05$; therefore, we are 95% confident to conclude that there is not enough evidence to support association between treatment group assignment and health status at randomization.

(b) Perform a longitudinal data analysis across all study follow-up visits (but not at randomization) to describe the relationship of the participants self-ratings as a function of the effects of health self-rating at the baseline, treatment group, month post randomization, and age group as predictors. Fit a GEE with unstructured correlation structure. Interpret your results.

```r
resp = subset(health.df1, time > "1")
resp$baseline = rep(subset(health.df1, time == "1")$health, as.numeric(table(resp$id)))

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

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

```
## running glm to get initial regression estimate
```

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

```r
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 + trt + time + agegroup, 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
## trtIntervention  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 log odds ratio of participants self-rating “good” vs. “poor” is 6.138, between participants self-rating “good” or “poor” at baseline, if take average among all measurements and all subjects within the same subgroup.
- The log odds ratio of participants self-rating “good” vs. “poor” is 8.162, between participants in “intervention” or “control” treatment group, if take average among all measurements and all subjects within the same subgroup.
- The log odds ratio of participants self-rating “good” vs. “poor” is 1.165, for per 3 months after randomization change, if take average among all measurements and all subjects within the same subgroup.
- The log odds ratio of participants self-rating “good” vs. “poor” is 3.861, between 25-34 age group vs. 15-24 age group, if take average among all measurements and all subjects within the same subgroup.
- The log odds ratio of participants self-rating “good” vs. “poor” is 3.861, between 35+ age group vs. 15-24 age group, if take average among all measurements and all subjects within the same subgroup.

(c) Fit a generalized linear mixed effects model with subject-specific random intercepts. Interpret your estimates. How are the interpretations different from the GEE model?

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

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: health ~ baseline + trt + 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 **
## trtIntervention   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 trtInt time   a25-34
## baseline    -0.526
## trtIntrvntn -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
```

Interpretation:

- The log odds ratio of participants self-rating “good” vs. “poor” is 1.224, for per 3 months after randomization change, if take average among all measurements and all subjects within the same subgroup.
- 2.781 is the average (conditional) log odds ratio of any paired subjects who only differ by the baseline self-reporting health status “good” or “poor”.

- 3.423 is the average (conditional) log odds ratio of any paired subjects who only differ by the treatment group (intervention vs. control).
- 2.25 is the average (conditional) log odds ratio of any paired subjects who only differ by the age group (age group 25-34 vs. 15-24)
- 2.25 is the average (conditional) log odds ratio of any paired subjects who only differ by the age group (age group 35+ vs. 15-24)

Difference between the two models interpretations:

GLMM model interpret the parameter as population average. In this case, it only interpret the time variable as within subject change, while interpret other variables as between subject change.