# P8131 HW8

Renjie Wei rw2844

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.

## Problem 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.

### Answers

Table 1: Logistic Model Coefficients

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|---|---|---|---|---|---|---|
| (Intercept) | -0.049 | 0.312 | -0.156 | 0.876 | -0.668 | 0.567 |
| TRTIntervention | -0.314 | 0.451 | -0.696 | 0.486 | -1.208 | 0.568 |

The odds ratio of self-reporting "good" health for status intervention group vs. control group at randomization (baseline) is 0.73. 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.

## Problem 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.

### Answers

I transformed the `TIME` to its original scale, that is $0, 3, 6, 12$.

The GEE model we are considering is

$$\log\{\frac{\Pr(Y_{ij}=1|X_{ij})}{\Pr(Y_{ij}=0|X_{ij})}\} = \log(\mu_{ij}) = \beta_0 + \beta_1 \times \text{BASELINEGood}_i + \beta_2 \times \text{Intervention}_i + \beta_3 \times \text{TIME}_i + \beta_4 \times \text{AGE:25-34}_i + \beta_5 \times \text{AGE:35}$$

since the study design is unbalanced plus the response is binary

$$\text{Var}(Y_{ij}|X_{ij}) = \phi v(\mu_{ij}) = \mu_{ij}(1-\mu_{ij})$$

finally, the within-subject association among the vector of repeated responses is assumed to have an unstructured pairwise log odds ratio pattern,

$$\log \text{OR}(Y_{ij}, Y_{ik}|X_{ij}, X_{ik}) = \alpha_{jk}$$

Table 2: Parameter estimates and standard errors from GEE model

| | Estimate | Naive S.E. | Naive z | Robust S.E. | Robust z |
|---|---|---|---|---|---|
| (Intercept) | -1.690 | 0.500 | -3.380 | 0.523 | -3.230 |
| BASELINE | 1.814 | 0.490 | 3.706 | 0.510 | 3.560 |
| TRTIntervention | 2.102 | 0.488 | 4.310 | 0.538 | 3.909 |
| TIME | 0.032 | 0.037 | 0.885 | 0.048 | 0.682 |
| AGEGROUP25-34 | 1.353 | 0.481 | 2.810 | 0.504 | 2.682 |
| AGEGROUP35+ | 1.421 | 0.798 | 1.781 | 0.784 | 1.813 |

Table 3: Working correlation estimates from GEE model

| obs | 2 | 3 | 4 |
|---|---|---|---|
| 2 | 1.000 | 0.172 | 0.586 |
| 3 | 0.172 | 1.000 | 0.201 |
| 4 | 0.586 | 0.201 | 1.000 |

**Interpretations:**

- The odds ratio of participants self-rating "good" vs. "poor" is 6.136, between participants self-rating "good" or "poor" at baseline, if take average among all measurements and all subjects within the same subgroup.
- The odds ratio of participants self-rating "good" vs. "poor" is 8.185, between participants in "intervention" or "control" treatment group, if take average among all measurements and all subjects within the same subgroup.

- The odds ratio of participants self-rating “good” vs. “poor” is 1.033, for per month after randomization change, if take average among all measurements and all subjects within the same subgroup.
- The odds ratio of participants self-rating “good” vs. “poor” is 3.867, between 25-34 age group vs. 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” vs. “poor” is 4.139, between 35+ age group vs. 15-24 age group, if take average among all measurements and all subjects within the same subgroup.

## Problem 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?

### Answers

The GLMM model we are considering is

$$\log\{\frac{\Pr(Y_{ij}=1|b_i)}{\Pr(Y_{ij}=0|b_i)}\} = \beta_0+\beta_1\times\text{BASELINEGood}_{ij}+\beta_2\times\text{Intervention}_{ij}+\beta_3\times\text{TIME}_{ij}+\beta_4\times\text{AGE:25-34}_{ij}+\beta_5\times\text{AGE:35+}_{ij}+b_i$$

Table 4: Estimated fixed effect for random intercept model

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -2.581 | 1.042 | -2.477 | 0.013 |
| BASELINE | 2.776 | 0.984 | 2.822 | 0.005 |
| TRTIntervention | 3.413 | 1.073 | 3.182 | 0.001 |
| TIME | 0.037 | 0.069 | 0.536 | 0.592 |
| AGEGROUP25-34 | 2.257 | 1.009 | 2.237 | 0.025 |
| AGEGROUP35+ | 1.982 | 1.381 | 1.435 | 0.151 |

Table 5: Covariance matrix for the estimates of fixed effects

| | (Intercept) | BASELINE | TRTIntervention | TIME | AGEGROUP25-34 | AGEGROUP35+ |
|---|---|---|---|---|---|---|
| (Intercept) | 1.085 | -0.648 | -0.712 | -0.030 | -0.656 | -0.608 |
| BASELINE | -0.648 | 0.968 | 0.473 | 0.001 | 0.376 | 0.373 |
| TRTIntervention | -0.712 | 0.473 | 1.151 | 0.004 | 0.427 | 0.305 |
| TIME | -0.030 | 0.001 | 0.004 | 0.005 | 0.001 | -0.001 |
| AGEGROUP25-34 | -0.656 | 0.376 | 0.427 | 0.001 | 1.018 | 0.543 |
| AGEGROUP35+ | -0.608 | 0.373 | 0.305 | -0.001 | 0.543 | 1.908 |

Table 6: Estimated random intercept

| Group | Variable | vcov | sd |
|---|---|---|---|
| ID | (Intercept) | 5.721 | 2.392 |

**Interpretation:**

- For the same subject, the odds ratio of participants self-rating "good" vs. "poor" is 1.038, for per month after randomization change, while holding other covariates fixed.
- For the same subject, The odds ratio of participants self-rating "good" vs. "poor" is 16.056, for participants self-rating "good" vs, "poor" at baseline, while holding other covariates fixed.
- For the same subject, The odds ratio of participants self-rating "good" vs. "poor" is 30.363, for participants in "intervention" vs. "control" treatment group, while holding other covariates fixed.

- For the same subject, The odds ratio of participants self-rating "good" vs. "poor" is 9.55, for participants in 25-34 age group vs. 15-24 age group, while holding other covariates fixed.
- For the same subject, The odds ratio of participants self-rating "good" vs. "poor" is 7.259, for participants in 35+ age group vs. 15-24 age group, while holding other covariates fixed.

**Difference between the GEE and GLMM parameter interpretations:**

- GEE is a marginal model focus on inferences about the study population.
- GLMM focus on inference about individuals.
- The parameters in GLMM measure the change in expected value of response while holding other covariates and *random effect* constant. On the contrary, the parameters in GEE measures the change in expected value of response, for *sub-populations* defined by fixed and known covariate values.

## Appendix

```
knitr::opts_chunk$set(echo = F,
                      message = F,
                      warning = F,
                      comment = "")
library(tidyverse)
library(readxl)
library(gee)
library(lme4)
library(nlme)
health.df = read_excel("./HW8-HEALTH.xlsx") %>%
  rename(TRT = TXT) %>%
  mutate(TRT = as.factor(TRT),
         HEALTH = as.numeric(HEALTH == "Good"),
         TIME = as.integer(TIME))
randomized.df = health.df %>% filter(TIME == 1)
randomized.fit = glm(HEALTH ~ TRT,
                  data = randomized.df,
                  family = binomial(link = "logit"))
randomized.fit %>% broom::tidy(conf.int = T) %>% knitr::kable(digits = 3, caption = "Logistic Model Coe
#randomized.fit %>% broom::glance() %>% knitr::kable(caption = "Logistic Model Goodness-of-Fit statisti
# remove the participants with randomization only, which is 108 and 115
health.lda = health.df %>%
    filter(!ID %in% names(which(table(health.df$ID) == 1)))
health.baseline = subset(health.lda, TIME == "1")
health.lda$BASELINE = rep(health.baseline$HEALTH, as.numeric(table(health.lda$ID)))
health.lda$TIME = health.lda$TIME - 1
health.lda = subset(health.lda, TIME > 0)
health.lda$TIME = case_when(health.lda$TIME == 1 ~ 3,
                            health.lda$TIME == 2 ~ 6,
                            health.lda$TIME == 3 ~ 12)

gee.fit = gee(HEALTH ~ BASELINE + TRT + TIME + AGEGROUP,
              data = health.lda,
              id = ID,
              family = "binomial",
```

```
              corstr = "unstructured",
              scale.fix = TRUE,
              scale.value = 1,

              )

gee.sum = summary(gee.fit)
gee.sum$coefficients %>% knitr::kable(digits = 3, caption = "Parameter estimates and standard errors fr
data.frame(gee.sum$working.correlation) %>% mutate(obs = c("2", "3","4")) %>% select(obs,everything()) %
glmm.fit = glmer(HEALTH ~ BASELINE + TRT + TIME + AGEGROUP + (1 | ID),
             data = health.lda,
             family = binomial)
glmm.sum = summary(glmm.fit)
glmm.sum$coefficients %>% knitr::kable(digits = 3, caption = "Estimated fixed effect for random intercep
as.data.frame(as.matrix(glmm.sum$vcov)) %>% knitr::kable(digits = 3, caption = "Covariance matrix for th
as.data.frame(glmm.sum$varcor) %>% select(-var2) %>% knitr::kable(digits = 3, caption = "Estimated rando
```