HW8_answer

Guojing Wu

4/22/2019

Problem

a) Cross-sectional relationship

We first delete some subjects that only have one observation: 108, 115. Then transform the baseline health self-rating to another column:

ID	TIME	TXT	HEALTH	AGEGROUP	baseline
101	2	Intervention	1	15-24	1
101	3	Intervention	1	15-24	1
101	4	Intervention	1	15-24	1
102	2	Control	0	15-24	0
102	3	Control	0	15-24	0
102	4	Control	0	15-24	0

We use GLM to build two models:

$$model1 : HEALTH \sim TIME + TXT * baseline + AGEGROUP$$
$$model2 : HEALTH \sim TIME + TXT + baseline + AGEGROUP$$

And used ANOVA to do deviance analysis to test whether the interaction term is significant or not. We got the pvalue = 0.1843, which state that we failed to reject the null hypothesis and the smaller model is better (in this case, model2).

b) GEE with unstructured correlation

Based on the question, the model here is $HEALTH \sim TIME + TXT + baseline + AGEGROUP$. We used GEE with unstructured correlation and get the parameters estimations:

	x
(Intercept)	-1.9220068
TIME	0.1530083
TXTIntervention	2.0995031
baseline	1.8144864
AGEGROUP25-34	1.3509848
AGEGROUP35+	1.4116600

Interpretation:

- the log odds ratio of being "Good" against "Poor" at self-rating is 0.153, for per 3 months change in time, if take average among all measurements and all subjects within the same subgroup (which is defined as share the same treatment, baseline and age group).

- the log odds ratio of being "Good" against "Poor" at self-rating is 2.1, between treatment and control, if take average among all measurements and all subjects within the same subgroup (which is defined as share the same time, baseline and age group).
- the log odds ratio of being "Good" against "Poor" at self-rating is 1.814, between being "Good" or "Poor" at the baseline, if take average among all measurements and all subjects within the same subgroup (which is defined as share the same time, treatment and age group).
- the log odds ratio of being "Good" against "Poor" at self-rating is 1.351, between "age group 25-34" and "age group 15-24", if take average among all measurements and all subjects within the same subgroup (which is defined as share the same time, treatment and baseline).
- the log odds ratio of being "Good" against "Poor" at self-rating is 1.412, between "age group 35+" and "age group 15-24", if take average among all measurements and all subjects within the same subgroup (which is defined as share the same time, treatment and baseline).

And the unstructured correlation matrix looks like below, the correlation between different times within the same subject varies.

1.0000000	0.1743007	0.5809889
0.1743007	1.0000000	0.2049833
0.5809889	0.2049833	1.0000000

c) GLMM with subject-specific random intercepts

Based on the question, the model here is $logitE(Y_{ij}|b_i) = (\beta_1 + b_{1i}) + \beta_2 TIME_{ij} + \beta_3 TXT_i + \beta_4 baseline_i + \beta_5 AGEGROUP_i$.

	x
(Intercept)	-2.9240338
TIME	0.2021495
TXTIntervention	3.4231159
baseline	2.7812900
AGEGROUP25-34	2.2587471
AGEGROUP35+	1.9802743

Interpretation:

- the log odds ratio of being "Good" against "Poor" at self-rating is 0.202, for 3 months change in time for the same subject.

The interpretation difference between GEE and GLMM is that:

- GEE interpret the parameters as population average
- GLMM interpret the parameters as subject-specific

Code

```
knitr::opts_chunk$set(echo = F,
                      message = F,
                      warning = F,
                      comment = "")
library(tidyverse)
```

```r
library(gee)
library(lme4)
library(nlme)
theme_set(theme_bw())
# original data
data.health <- readxl::read_xlsx("HW8-HEALTH.xlsx") %>%
  mutate(TXT = as.factor(TXT),
         HEALTH = as.numeric(HEALTH == "Good"),
         TIME = as.integer(TIME))
# remove subjects with only one obs
data.hea1 <- data.health %>%
  filter(!ID %in% names(which(table(data.health$ID) == 1))) # remove the subject only has baseline
# transform baseline value to another column
data.hea2 <- data.hea1 %>% filter(TIME != 1)
data.hea2$baseline = rep(subset(data.hea1, TIME == "1")$HEALTH, as.numeric(table(data.hea2$ID))) # crea

# GLM to test cross-sectional relationship
hea_glm1 = glm(HEALTH ~ TIME + TXT * baseline + AGEGROUP, data = data.hea2)
hea_glm2 = glm(HEALTH ~ TIME + TXT + baseline + AGEGROUP, data = data.hea2)
ano = anova(hea_glm2, hea_glm1)
pvalue = 1 - pchisq(ano$Deviance[2], df = ano$Df[2])

head(data.hea2) %>% knitr::kable()
# GEE with unstructured correlation
hea_gee1 = gee(HEALTH ~ TIME + TXT + baseline + AGEGROUP, data = data.hea2, family = "binomial", id = I
hea_gee1$coefficients %>% knitr::kable()
hea_gee1$working.correlation %>% knitr::kable()
# GLMM with subject-specific random intercepts
hea_glmer = glmer(HEALTH ~ TIME + TXT + baseline + AGEGROUP + (1 | ID), data = data.hea2, family = "bin
fixed.effects(hea_glmer) %>% knitr::kable()
```