hw8-Generalized Linear models for longitudinal Data case study

Cary Ni

```
# load dataset and modify variable types
data_df = readxl::read_excel("./data/HW8-HEALTH.xlsx") %>% janitor::clean_names()
data_df[, c(3:5)]= lapply(data_df[, c(3:5)], as.factor)
summary(data_df)
```

```
##        id           time                txt       health     agegroup
##  Min.   :101   Min.   :1.00   Control     :141   Good:174   15-24:127
##  1st Qu.:120   1st Qu.:1.00   Intervention:138   Poor:105   25-34:126
##  Median :140   Median :2.00                                 35+  : 26
##  Mean   :287   Mean   :2.33
##  3rd Qu.:605   3rd Qu.:3.00
##  Max.   :625   Max.   :4.00
```

```
# data wrangling
# 1 == good
data_base = subset(data_df, time == 1) %>%
  mutate(baseline = health) %>%
  mutate(n_health = as.numeric(health == "Good"))

data_sub = data_base %>% select(id, baseline)

data_new = data_df %>%
  left_join(data_sub, by = "id") %>%
  filter(time != 1) %>%
  mutate(n_health = as.numeric(health == "Good"))
```

(a)

```
# fit cross-sectional model at randomization
health_glm = glm(n_health ~ txt, data = data_base, family = "binomial")
summary(health_glm)
```

```
##
## Call:
## glm(formula = n_health ~ txt, family = "binomial", data = data_base)
##
## 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
## 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
```

At the time of randomization, it can be seen that odds of being good in `health` is lower for the intervention group compared to control group with a factor of 0.731. However, the p value gives 0.486, meaning that this difference between two groups is not significant at 0.05 significance level. We can say that the baseline odds of being good in `health` is not significantly different for intervention group and control group.

(b)

```
# fit gee after randomization
health_gee = gee(n_health ~ baseline + txt + time + agegroup, data = data_new,
                 family = "binomial", id = id, corstr = "unstructured", scale.fix = FALSE)
```

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

## running glm to get initial regression estimate

##     (Intercept)    baselinePoor txtIntervention            time   agegroup25-34
##      -0.0301908      -1.7112931       1.9977806       0.1321222       1.1958638
##     agegroup35+
##       1.3954271
```

```
summary(health_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 = n_health ~ baseline + txt + time + agegroup, id = id,
##     data = data_new, 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)     -0.1075204  0.7564535 -0.1421375   0.7206791 -0.1491931
## baselinePoor    -1.8144864  0.6033350 -3.0074276   0.5104410 -3.5547427
## txtIntervention  2.0995031  0.6008738  3.4940832   0.5379270  3.9029513
## time             0.1530083  0.2017530  0.7583941   0.2107268  0.7260980
## agegroup25-34    1.3509848  0.5930043  2.2782040   0.5038608  2.6812659
## agegroup35+      1.4116600  0.9825238  1.4367693   0.7864438  1.7949916
##
## 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
```

From the GEE model with unstructured correlation, it can be seen from coefficients that the baseline of being poor decrease the odds of being good in level of health to a factor of 0.163 compared to being good in baseline while holding other covariates fixed. This term is also shown to be significant as its p value is less than 0.05 at the given significance level. Holding other covariates fixed, the intervention group is shown to be have a 8.166 times odds of reporting good in level of health compared to control group with a p value less than 0.05 also suggests its significance. Holding other covariates fixed, with every unit increase in time (3 months), the odds of reporting good increase to a factor of 1.165 though this term is shown to be insignificant with a p value larger than 0.05. Holding other covariates fixed, compared to age group (15-24), age group (25-34) has 3.857 times the odds of reporting good with significance in p value less than 0.05, whereas age group (35+) is not significantly different from the group (15-24) with a p value larger than 0.05 even though it has 4.104 the odds of reporting good compared to group (15-24).

(c)

```
# Fit the model with random intercept
health_glmm = glmer(n_health ~ baseline + txt + time + agegroup + (1 | id),
                    data = data_new,
                    family = "binomial")
summary(health_glmm)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: n_health ~ baseline + txt + time + agegroup + (1 | id)
##    Data: data_new
##
##      AIC      BIC   logLik deviance df.resid
##    184.8    207.9    -85.4    170.8      192
##
## Scaled residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -2.5390 -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)      -0.1428     1.1479  -0.124  0.90102
## baselinePoor     -2.7813     0.9873  -2.817  0.00485 **
## txtIntervention   3.4231     1.0779   3.176  0.00150 **
## time              0.2022     0.3090   0.654  0.51295
## agegroup25-34     2.2587     1.0128   2.230  0.02573 *
## agegroup35+       1.9803     1.3853   1.430  0.15285
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) bslnPr txtInt time   a25-34
## baselinePor -0.264
## txtIntrvntn -0.228 -0.450
## time        -0.742 -0.034  0.068
## agegrp25-34 -0.256 -0.380  0.396  0.022
## agegroup35+ -0.150 -0.275  0.206 -0.002  0.390
```

From the general linear mixed effects model, it can be seen from coefficients that the baseline of being poor decrease the odds of being good in level of health to a factor of 0.062 compared to being good in baseline while holding other covariates fixed. This term is also shown to be significant as its p value is less than 0.05 at the given significance level. Holding other covariates fixed, the intervention group is shown to be have a 30.569 times odds of reporting good in level of health compared to control group with a p value less than 0.05 also suggests its significance. Holding other covariates fixed, with every unit increase in time (3 months), the odds of reporting good increase to a factor of 1.224 though this term is shown to be insignificant with a p value larger than 0.05. Holding other covariates fixed, compared to age group (15-24), age group (25-34) has 9.583 times the odds of reporting good with significance in p value less than 0.05, whereas age group (35+) is not significantly different from the group (15-24) with a p value larger than 0.05 even though it has 7.243 the odds of reporting good compared to group (15-24).

It can seen that the coefficients in general linear mixed effects model differ from GEE model in the magnitude but not the significance. In GLLM model, `baseline`, `intervention`, `time`, `age groups` leads to greater change in the odds of reporting good compared to GEE model while maintaining the equivalent sign of all coefficients. On the other hand, the Z statistics of all coefficients in GLLM model are less than (in magnitude) the Z statistics of the corresponding coefficients in GEE model, meaning the estimated coefficients are of less confidence though the reported significance stay the same at 0.05 level.