p8131_hw8_xy2395

Jack Yan

4/18/2019

```
library(tidyverse)
library(readxl)
library(gee)
library(lme4)
health_df = read_xlsx('../hw8/HW8-HEALTH.xlsx')
```

```
# Data manipulation
health_df <-
  health_df %>%
  janitor::clean_names()
```

(a)

```
# Use baseline data only
health_baseline <-
  health_df %>%
  filter(time == 1)
```

```
# 2-way table
table(health_baseline$txt, health_baseline$health) %>%
  addmargins()  %>%
  knitr::kable()
```

	Good	Poor	Sum
Control	20	21	41
Intervention	16	23	39
Sum	36	44	80

```
# expected values
table(health_baseline$txt, health_baseline$health) %>%
  chisq.test() %>% .$expected  %>%
  knitr::kable()
```

	Good	Poor
Control	18.45	22.55
Intervention	17.55	21.45

```
# chi-squared test for association between assignment and health rating
table(health_baseline$txt, health_baseline$health) %>%
  chisq.test()
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
```

```
## data:  .
## X-squared = 0.22287, df = 1, p-value = 0.6369
```

We can see from the 2-way table that the number of people randomized to control group who rated their health status as 'Good' is 20, while its expected value is 18.45. The difference between observed and expected values is acceptable. The chi-squared test (p-value = 0.6369 >0.05) also suggests that evidence is not strong enough to conclude association between treatment group and health status at baseline.

Also use logistic regression to evaluate the relationship between treatment group and health self-rating at baseline, adjusting for age group.

```
health_baseline1 <-
   health_baseline %>%
   mutate(health = if_else(health == 'Good', 1, 0),
          health = as.factor(health))
# Logistic regression
fit_glm = glm(health ~ txt + agegroup,
              data = health_baseline1,
              family = binomial)
summary(fit_glm)
```

```
##
## Call:
## glm(formula = health ~ txt + agegroup, family = binomial, data = health_baseline1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2191  -1.0831  -0.8676   1.2058   1.6687
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.0976     0.3834   0.255    0.799
## txtIntervention  -0.3234     0.4554  -0.710    0.478
## agegroup25-34    -0.1643     0.4734  -0.347    0.729
## agegroup35+      -0.8808     0.8988  -0.980    0.327
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 110.10  on 79  degrees of freedom
## Residual deviance: 108.57  on 76  degrees of freedom
## AIC: 116.57
##
## Number of Fisher Scoring iterations: 4
```

```
exp(-0.3234)
```

```
## [1] 0.7236843
```

```
c(-0.3234-1.96*0.4554, -0.3234+1.96*0.4554) %>% exp()
```

```
## [1] 0.2964182 1.7668247
```

The estiamted odds ratio of Good health for Intervention group vs Control group is 0.724. However, the p-value for the coefficient is 0.478 >0.05, and the 95% confidence interval for the odds ratio is (0.296, 1.767), so we conclude that there is not enough evidence to support association between treatment group assignment and health status at baseline.

(b) GEE

```
health_new =
  health_baseline %>%
  rename(baseline = health) %>%
  select(id, baseline) %>%
  inner_join(., health_df, by = 'id') %>%
  filter(time != 1) %>%
  # recode `months` such that it reflects the number of months post randomization
  mutate(months = 3 * (time - 1),
         months = if_else(months == 9, 12, months),
         health = if_else(health == 'Good', 1, 0),
         baseline = fct_relevel(baseline, 'Poor'))

fit_gee =
  gee(health ~ baseline + txt + months + agegroup,
      id = id, scale.fix = TRUE, scale.value = 1,
      family = binomial,
      corstr = 'unstructured',
      data = health_new)
```

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

## running glm to get initial regression estimate

##     (Intercept)    baselineGood txtIntervention          months
##     -1.52535766      1.71063852      1.99669985      0.02536275
##   agegroup25-34     agegroup35+
##      1.19749448      1.39742621
```

```
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 = health ~ baseline + txt + months + agegroup, id = id,
##     data = health_new, family = binomial, corstr = "unstructured",
##     scale.fix = TRUE, scale.value = 1)
##
## Summary of Residuals:
##         Min          1Q      Median          3Q         Max
## -0.98144969 -0.18317233  0.08914345  0.17159228  0.83093959
##
##
## Coefficients:
##                    Estimate Naive S.E.    Naive z Robust S.E.   Robust z
## (Intercept)     -1.68960132 0.49985657 -3.3801723  0.52303338 -3.2303891
## baselineGood     1.81418056 0.48958528  3.7055456  0.50961334  3.5599158
```

```
## txtIntervention  2.10225898 0.48779381  4.3097286  0.53777951  3.9091467
## months           0.03243343 0.03665686  0.8847848  0.04755408  0.6820326
## agegroup25-34    1.35250468 0.48130172  2.8100973  0.50420159  2.6824681
## agegroup35+      1.42052166 0.79781620  1.7805124  0.78372968  1.8125148
##
## Estimated Scale Parameter:  1
## Number of Iterations:  5
##
## Working Correlation
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.1719328 0.5859907
## [2,] 0.1719328 1.0000000 0.2013998
## [3,] 0.5859907 0.2013998 1.0000000
```

```
# Confidence intervals
tibble(term = names(fit_gee$coef),
       coef = fit_gee$coef,
       std_err = sqrt(diag(fit_gee$robust.variance)),
       CIL = fit_gee$coef - 1.96 * std_err,
       CIR = fit_gee$coef + 1.96 * std_err,
       p_value = 2*pnorm(-abs(coef/std_err)))
```

```
## # A tibble: 6 x 6
##   term               coef std_err     CIL    CIR   p_value
##   <chr>             <dbl>   <dbl>   <dbl>  <dbl>     <dbl>
## 1 (Intercept)     -1.69    0.523  -2.71   -0.664 0.00124
## 2 baselineGood     1.81    0.510   0.815   2.81  0.000371
## 3 txtIntervention  2.10    0.538   1.05    3.16  0.0000926
## 4 months           0.0324  0.0476 -0.0608  0.126 0.495
## 5 agegroup25-34    1.35    0.504   0.364   2.34  0.00731
## 6 agegroup35+      1.42    0.784  -0.116   2.96  0.0699
```

```
fit_gee$coef / sqrt(diag(fit_gee$robust.variance))
```

```
##     (Intercept)    baselineGood txtIntervention          months
##      -3.2303891       3.5599158       3.9091467       0.6820326
##   agegroup25-34     agegroup35+
##       2.6824681       1.8125148
```

The log odds ratio of good status for people who rated themselves 'Good' vs 'Poor' at baseline is 1.81, among subpopulation with the same treatment group assignment, months post randomization, and age group. The 95% confidence interval for log odds ratio is (0.815, 2.81).

The log odds ratio of good status for people in Intervention vs Control group is 2.10, among subpopulation with the same baseline self-rating, months post randomization, and age group. The 95% confidence interval for log odds ratio is (1.05, 3.16).

The log odds ratio of good status for every additional month post randomization is 0.0324, among subpopulation with the same baseline self-rating, treatment group assignment, and age group. The 95% confidence interval for log odds ratio is (-0.06, 0.126).

The log odds ratio of good status for age group 25-34 vs age group 15-24 is 1.35, among subpopulation with the same baseline self-rating, months post randomization, and treatment. The 95% confidence interval for log odds ratio is (0.364, 2.34).

The log odds ratio of good status for age group 35+ vs age group 15-24 is 1.42, among subpopulation with the same baseline self-rating, months post randomization, and treatment. The 95% confidence interval for log odds ratio is (-0.116, 2.96).

(c) GLMM

```
fit_glmm =
  glmer(health ~ baseline + txt + months + agegroup + (1 | id),
        data = health_new,
        family = binomial)

summary(fit_glmm)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: health ~ baseline + txt + months + agegroup + (1 | id)
##    Data: health_new
##
##      AIC      BIC   logLik deviance df.resid
##    185.0    208.0    -85.5    171.0      192
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.6112 -0.2327  0.1402  0.2982  1.8239
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  id     (Intercept) 5.721    2.392
## Number of obs: 199, groups:  id, 78
##
## Fixed effects:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -2.58087    1.04177  -2.477  0.01323 *
## baselineGood     2.77609    0.98380   2.822  0.00478 **
## txtIntervention  3.41322    1.07266   3.182  0.00146 **
## months           0.03718    0.06933   0.536  0.59178
## agegroup25-34    2.25652    1.00877   2.237  0.02529 *
## agegroup35+      1.98229    1.38117   1.435  0.15122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) bslnGd txtInt months a25-34
## baselineGod -0.632
## txtIntrvntn -0.637  0.449
## months      -0.409  0.016  0.047
## agegrp25-34 -0.624  0.379  0.395  0.007
## agegroup35+ -0.422  0.274  0.206 -0.007  0.390
```

```
tibble(
  term = names(coef(summary(fit_glmm))[, 'Estimate']),
  coef = coef(summary(fit_glmm))[, 'Estimate'],
  se = sqrt(diag(vcov(fit_glmm))),
  CIL = coef - 1.96 * se,
  CIR = coef + 1.96 * se
)
```

```
## # A tibble: 6 x 5
##   term               coef     se     CIL    CIR
##   <chr>             <dbl>  <dbl>   <dbl>  <dbl>
## 1 (Intercept)     -2.58   1.04   -4.62   -0.539
## 2 baselineGood     2.78   0.984   0.848   4.70
## 3 txtIntervention  3.41   1.07    1.31    5.52
## 4 months           0.0372 0.0693 -0.0987  0.173
## 5 agegroup25-34    2.26   1.01    0.279   4.23
## 6 agegroup35+      1.98   1.38   -0.725   4.69
```

Different from GEE model, here for GLMM we can only interpret the `months` term, because other terms cannot be changed within the same subject during the course of the study. The interpretation of the coefficient for `months` is: the log odds ratio of 'Good' rating for one additional month post randomization is 0.0372, within the same subject. The 95% confidence interval is (-0.0987, 0.173), and p-value is 0.592 >0.05, so the `months` term is not significant in this model.